

QUANTITATIVE IMAGING BIOMARKERS: COMBINING DATA-CENTRIC DEEP
LEARNING WITH ANATOMICAL CONTEXT

By

Yucheng Tang

Dissertation

Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in

Electrical and Computer Engineering

August 12, 2022

Nashville, Tennessee

Approved:

Bennett Landman, Ph.D.

Yuankai Huo, Ph.D.

Richard G. Abramson, M.D., M.S.

Thomas Lasko, M.D-Ph.D.

Zhoubing Xu, Ph.D.

Copyright © 2022 Yucheng Tang
All Rights Reserved

ACKNOWLEDGMENTS

My fateful journey as a graduate student with MASI lab and Vanderbilt has been a personal and collaborative adventure. First, I am grateful and indebted to my fiancée and parents for their support, love and understanding. They encourage me to dive into the world of science and research. My mom and dad, Shi-Xian Li and He-Ping Tang, you care about my livings 7000 miles away, you've always supported me, made me happy and keep me aware I have your back in this big world. You encourage me to see the colorful world and to pursue my dream of science and research. You showed me the magic world of science and engineering and opened the door for experiments and critical thinking. Xu-Juan Sun, my girl, thank you for always supporting me, my heartfelt gratitude goes out with you no matter we are in same place or in different cities, countries. I feel like I am the most lucky man with you. Haocheng Tang, my little brother, you are smart. I remember the time when you were choosing majors, you love biology and chemistry, but I recommend you the medicine and to pursue M.D. for your life, I know it's hard and hope you don't hate me.

I'm deeply appreciative to the opportunities at Siemens Healthineer before I started my research career. My mentor S. Kevin Zhou and G. Funka-Lea brought me into the fascinating world of research. I never felt attracted and settled to devoted career for my life until the three months at Princeton, New Jersey. Professor S. Kevin Zhou and many friends, colleagues, they taught me to be rigorous about science and research, be humble, and young for new techniques in a life-long journey. In particular, with Haofu Liao, I learned the important skills of deep learning for my first project on cardiac MRI. Dong Yang, Zizhao Zhang, He Zhang - it's been great being with you in a room full of ideas and passions, you're the best young generation of your research area. Thanks for being patient with me as I was a completely new to machine learning, medical imaging and research.

I'm so thankful for having Yuankai Huo, Zhoubing Xu as my mentor and friends from Siemens. You are perfect scientist that showed me the veracity, character and morality of being a scientist. It's you that inspired me to apply to Vanderbilt and join MASI Group, and I never regret to accept the Vandy offer immediately after I got the opportunity. I learned as much from you as scientist of

medical image analysis and computer vision, the joys of statistical methods and machine learning attract me to investigate the innovative approaches for healthcare. Dr. Yunqiang Chen and Dr. Dashan Gao from 12 Sigma, I feel so fortunate to work with you as you are one of the best scientists that make research into solid products.

I would not be the scientist today without the warm and open environments at MASI Lab, Vanderbilt. The MASI Lab is my first real home at a different country, and MASI Lab represent not only the laboratory but the family that support everyone. Dr. Yuankai Huo and Dr. Shunxing Bao, thank you for being my mentor and fountain of inspirations and advice that introduce me to the academic thinking. Riqiang Gao and Vishwesh Nath, we spent a lot of night together at Featheringill hall, and talked about our future in 2018-2019, it's fortunate grow with you. Vishwesh Nath, thank you for being my mentor during the internship at Nvidia, I can't imagine what's the days would be without you. Cam, Colin, Karthik thank you for showing me the excitement life and the advice of communication in the MASI Lab; Leon, Lucas, Nancy, Praitayini, Thomas, Qi, Kaiwen, you are great researchers and colleagues that make MASI lab as a family. Cailey, you've been a pilot for making happiness for our MASI family; Peter and Xin, it's so lucky to work with you on the same body group, you are very positive and making our projects vigorous. My undergraduates and interns, Yiyuan Yang, Olivia, Yuchen, Yan, Jason Can, and Yinchu Zhou, you are excellent students and I learned much from your vitality and passion.

I feel grateful to our collaborators from Vanderbilt University Medical Center, professor Richard Abramson, Michael Savona, Fogo Agnes, Jeffrey Spraggins, Jeff Carr and Thomas Lasko, I'm obsessed to these clinical perspectives of science and technology. Our research ideas come from the real clinical useful problems and make our contributions practical in daily use. Professor Bennett Landman, you are the courageous and visionary leader, advisor and friend. You have shown me the most passion and morality for being a scientist, I learned the enthusiasm, the vitality, the trustworthy, the determination and creativity for the leadership. I'm so honored and deeply indebted. Thank you.

Finally, I feel fortunate and grateful to be a scientist in the medical image analysis community around the world. Our efforts can make a difference to human health and make a better world with at least small milestones.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	iii
LIST OF TABLES	xiii
LIST OF FIGURES	xvi
1 Introduction	1
1.1 Overview	1
1.2 Abdominal Imaging	2
1.2.1 Abdomen CT Field of Views (FOVs)	3
1.2.2 CT Enhancement Phases	4
1.2.3 CT Resolution	5
1.3 Body Compositions	5
1.3.1 Abdominal Organs	5
1.3.2 Body parts	8
1.4 Large Scale Medical Data	8
1.5 Machine Learning in Medical Image Analysis	9
1.5.1 Medical Image Segmentation	10
1.5.2 Self-Supervised Learning	10
1.5.3 Transformers	11
1.6 Multi-Modal Clinical Context	13
1.6.1 Predictive Modeling with EHR	13
1.6.2 Multi-modal Representation Learning for Medical Data	14
1.7 Challenges in Modern Medical Image Analysis	15
1.7.1 Variations in Medical Image FOV	15

1.7.2	CT Enhancement Phase Identification	15
1.7.3	High-Resolution 3D Medical Image Segmentation	16
1.7.4	Body Reference Space	16
1.7.5	Combining Clinical Context	17
1.7.6	Modeling Medical Spatial Representations	18
1.8	Contributed Work	18
1.8.1	Contribution 1: Body Part Regression for CT	18
1.8.2	Contribution 2: CT Contrast Phase Identification	19
1.8.3	Contribution 3: High-Resolution 3D Medical Image Segmentation	20
1.8.4	Contribution 4: Combining Non-Imaging Clinical Context	20
1.8.5	Contribution 5: Spatial Long-range Dependencies	20
1.8.6	Contribution 6: Efficient 3D Transformer Models with Self-Supervised Learning	21
1.8.7	Contribution 7: Construction of 3D abdominal atlases and tissue corre- spondence modeling	21
1.8.8	Contribution 8: Application of Quantitative Imaging Biomarkers	22
2	Body Part Regression with Self-Supervision	23
2.1	Introduction	23
2.2	Data	25
2.3	Experiment	26
2.3.1	Body Part Regression	26
2.3.2	Robust Regression Refinement	27
2.3.3	Deep Supervision Network	27
2.4	Results	28
2.5	Discussion	31
3	Anatomy-Aware Semi-Supervised Body Part Regression for Multi-Contrast CT Im- ages	33
3.1	Introduction	33

3.2	Methods	35
3.2.1	Localization by Anatomy Landmarks	35
3.2.2	Multi-Contrast Dataset Selection	35
3.2.3	Data Pre-Processing	36
3.2.4	Transformer-based MixMatch Model	36
3.3	Experiments	39
3.4	Results and Discussion	40
3.4.1	Anatomies landmark boosted regression accuracy	40
3.4.2	Robustness to Multi-Contrast CT Scans	42
3.4.3	Reproducibility and Repeatability	42
3.5	Conclusion	42
4	Phase Identification for Dynamic CT enhancements with Generative Adversarial Network	43
4.1	Introduction	43
4.2	Materials and Methods	44
4.2.1	Dataset of Studies	44
4.2.2	Preprocessing	46
4.2.3	Contrast Disentangling Generative Adversarial Network (CD-GAN)	46
4.2.4	Baseline Architectures	49
4.2.5	Training Parameters	51
4.2.6	Experimental Design	51
4.2.7	Evaluation Metrics	52
4.3	RESULTS	52
4.3.1	Quantitative Evaluation	52
4.3.2	Qualitative Evaluation	54
4.4	Discussion	56
4.5	Conclusion	58
5	High-resolution 3D Abdominal Segmentation with Random Patch Network Fusion	59

5.1	Introduction	59
5.2	Theory	62
5.2.1	Stage 1: Preliminary segmentation	64
5.2.2	Stage 2: Random patch sampling	64
5.2.3	Stage 3: Label fusion	65
5.3	Methods	65
5.3.1	Preprocessing and body part regression	67
5.3.2	Baseline architectures	67
5.3.3	Implementation Details	68
5.3.4	Experimental Design	69
5.3.4.1	Random Patch Network Fusion	70
5.3.4.2	Ablation Study	70
5.3.5	Validation on External Datasets	71
5.3.6	Evaluation Metrics	72
5.4	Results	72
5.4.1	Random Patch Network Fusion	72
5.4.2	Ablation Study	78
5.4.2.1	Effect of Patch-Based Strategies	78
5.4.2.2	Effect of Average Number of Coverages per Voxel	79
5.4.2.3	Effect of Patch Size	80
5.4.2.4	Validation on External Datasets	80
5.4.3	Comparison of State-of-the-art Methods	81
5.4.3.1	Coarse-to-fine Methods	81
5.4.3.2	Patch Selection Methods	83
5.4.3.3	Comparison of Time Efficiencies with Different Methods	83
5.4.3.4	Comparison with Different Medical Image Segmentation Methods	84
5.4.3.5	Comparison with Multi-Atlas Abdomen Labeling Challenge leader-board	84
5.5	Conclusion and Discussion	85

6	Pancreas CT Segmentation by Predictive Phenotyping	87
6.1	Introduction	87
6.2	Method	88
6.2.1	Problem Formulation	88
6.2.2	Loss Functions	90
6.2.3	Phenotype Embedding	90
6.3	Experiments	91
6.3.1	Dataset	91
6.3.2	Implementation Details	91
6.3.3	Comparison with State-Of-The-Arts	92
6.3.4	Results	92
6.4	Discussion and Conclusion	95
7	Spatial Long-Range Dependencies: Transformers for 3D Medical Image Segmentation	96
7.1	Introduction	96
7.2	Data	96
7.3	Experiment	98
7.4	Results	99
7.5	Discussion	99
8	Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis	101
8.1	Introduction	101
8.2	Related Works	104
8.2.1	Medical Segmentation with Transformers	104
8.2.2	Pre-training in Medical Image Analysis	105
8.3	Swin UNETR	105
8.3.1	Swin Transformer Encoder	105
8.3.2	Decoder	107

8.4	Pre-training	107
8.4.1	Masked Volume Inpainting	107
8.4.2	Image Rotation	108
8.4.3	Contrastive Coding	109
8.4.4	Loss Function	109
8.5	Experiments	109
8.5.1	Datasets	109
8.5.2	Implementation Details	111
8.5.3	Evaluation Metrics	111
8.5.4	Results	112
8.5.4.1	BTCV Multi-organ Segmentation Challenge	112
8.5.4.2	Segmentation Results on MSD	113
8.5.5	Ablation Study	113
8.5.5.1	Efficacy of Pre-training	113
8.5.5.2	Reduce Manual Labeling Efforts	114
8.5.5.3	Size of Pre-training Dataset	115
8.5.5.4	Efficacy of Self-Supervised Objectives	115
8.6	Discussion and Limitations	115
8.7	Conclusions	116
8.8	Appendix	117
8.9	Pre-training Datasets	117
8.10	Preprocessing Pipelines	118
8.10.1	BTCV Dataset	118
8.10.2	MSD Dataset	120
8.11	Results	122
8.11.1	MSD Qualitative Comparisons	122
8.11.2	MSD Quantitative Comparisons	123
8.12	Model Complexity and Pre-training Time	123
8.13	Pre-Training Algorithm Details	124

9	Characterizing Renal Structures with 3D Block Aggregate Transformers	125
9.1	Introduction	125
9.2	Related Works	126
9.3	Method	127
9.3.1	UNesT Architecture	127
9.3.2	3D Block Aggregation	129
9.3.3	Decoder	129
9.4	Experiments	129
9.4.1	Dataset	129
9.4.2	Implementation Details	130
9.5	Results	131
9.5.1	Characterization of Renal Structures	131
9.5.2	Ablation Study	131
9.6	Discussion and Conclusion	133
10	Abdominal Atlas Template for Accurate Tissue Correspondence	135
10.1	Introduction	135
10.2	Data	137
10.3	Experiments	139
10.3.1	Kidney and Pancreas Atlas	139
10.3.2	Automated Renal Structures Segmentation	141
10.4	Results	142
10.5	Conclusion	147
11	Clinical Application: Validation and Estimation of Spleen Volume Via Computer-assisted Segmentation on Clinically Acquired CT Scans	149
11.1	Introduction	149
11.2	Materials and Methods	151
11.3	Analysis.	154

11.4	Results	155
11.5	Discussion	156
11.6	Clinical Improvement	157
11.7	Summary	157
12	Conclusion and Future Works	158
12.1	Impact of the Dissertation	158
12.2	Visions Beyond Medical Image Analysis	160
12.3	Future Works	161
12.3.1	Scaling Transformer Medical Segmentation Models for Large-Scale Data	161
12.3.2	Understanding Imaging Biomarks for Type 1 Diabetes with CT and MRI images	161
12.3.3	Learning Continuously from Incremental Data and Organ Segmentation .	162
12.3.4	Collaborative Learning with Federated Clients for Healthcare	162
	References	163
A	Copyright from Publishers	185
A.1	Copyright from arXiv	185
A.2	Copyright from Elsevier	186
A.3	Copyright from LNCS	187
A.4	Copyright from SPIE and JMI	188
A.5	Copyright from IEEE Conference and IEEE TMI	189
A.6	Copyright from Medical Physics	189

LIST OF TABLES

Table		Page
2.1	Multi-organ segmentation results with 3D U-Net and different preprocessing strategies are presented with average Dice coefficients. The best performance results marked as bold. BR means bodypart regression. “*” indicates statistically significant (p-value < 0.01 paired t-test) between left and right mean DSC.	31
3.1	Definition of anatomy-based landmarks for body parts. We demonstrate multiple landmarks in abdomen regions for diverse organs such as liver, pancreas, kidneys.	35
3.2	The quantitative results comparison with baseline methods. The absolute error and scale percentage error are shown according to the 8 manual defined landmarks.	39
3.3	The testing results categorized by different CT contrast phases. The performance show late arterial and portal venous phase data are of higher accuracy partially due to the larger size of training data.	41
3.4	The results show the reproducibility and repeatability of the proposed automatic method and two independent readers. It demonstrate the error between SemiBR and readers are similar to inter-reader assessments.	41
4.1	Quantitative comparison of scan-level performance. We mark * to indicate statistically significant (p-value < 0.01, Stuart-Maxwell test) compared to previous method. Best values are marked bold.	55
5.1	Mean DSC and variance of 12 abdominal organs compared with our method and three baseline approaches on BTCV miccai2015 challenge testing cohort. Our method presented significant improvement compared to Hierarchical method. (p-value < 0.01 with paired t-test). Note: Bold values indicates best mean DSC of each organ.	73

5.2	Segmentation performance of models trained on BTCV dataset in Mean DSC and variance, tested on HEM1538 and ImageVU pancreas, the proposed method is compared with baselines (p-value < 0.01 with paired t-test between random patch network fusion and two-level hierarchy).	73
5.3	Comparison of coarse-to-fine methods between our proposed approach and state-of-the-art methods. The evaluation is conducted on BTCV testing dataset in terms of mean DSC.	76
5.4	Fine stage performance comparison with state-of-the-art methods on patch selection strategies using same backbone network (3D UNet). The evaluation is performed on BTCV testing data on 12 abdominal organs in terms of mean and variance.	77
5.5	Average time cost per CT volume in the testing phase on different multi-organ segmentation models, where mean DSC is the average Dice score across 12 organs on BTCV testing data.	79
5.6	Evaluation of different medical image segmentation methods on the BTCV testing dataset in multi-organ segmentation (12 organs). The evaluation is performed in terms of mean DSC and across organs.	81
5.7	Leaderboard of Multi-Atlas Abdomen Labeling Challenge (mean DSC).	81
6.1	Performance comparison on the abnormal pancreas segmentation dataset. C2F denotes coarse-to-fine training. * denotes statistically significant against above method with Wilcoxon signed-rank test.	93
6.2	External testing performance comparison on BTCV MICCAI Challenge 2015 and TCIA pancreas (mean DSC) with our model trained on the internal data. Note that no subject from these two datasets are used for training. C2F denotes coarse-to-fine training strategies. * for statistically significant against above method with Wilcoxon signed-rank test.	93
8.1	caption	110
8.2	Overall performance of top-ranking methods on all 10 segmentation tasks in the MSD public test leaderboard. NSD denotes Normalized Surface Distance. . . .	112

8.3	MSD test dataset performance comparison of Dice and NSD. Benchmarks obtained from MSD test leaderboard	112
8.4	MSD test dataset performance comparison of Dice and NSD. Benchmarks obtained from MSD test leaderboard ¹	112
8.5	Ablation study of the effectiveness of each objective function in the proposed pre-training loss. HD denotes Hausdorff Distance. Experiments on fine-tuning the BTCV dataset.	114
8.6	Summary of datasets for pre-training, the use of cohorts identifies diversified regions of interest.	120
8.7	Additional MSD MRI test dataset performance comparison of Dice and NSD. Benchmarks obtained from MSD test leaderboard. Task01 BrainTumour results are shown in the paper. Note: The results reported for TransVW [1] and Models Genesis [2] are from the official leaderboard for MRI tasks.	121
8.8	Comparison of number of parameters, FLOPs and averaged inference time for various models in BTCV experiments.	121
9.1	Segmentation results of the renal substructure on testing cases. The UNesT achieves state-of-the-art performance compared to prior kidney components studies and 3D medical segmentation baselines. The number of parameters and GFLOPS (with a single input volume of $96 \times 96 \times 96$) are shown for deep learning-based approaches. * indicates statistically significant ($p < 0.01$) by Wilcoxon signed-rank test.	130
9.2	Comparison of volumetric analysis metrics between the proposed method and the state-of-the-art clinical study on kidney components.	131
10.1	Summarized statistics for the final automatic system compared to manual segmentation.	142
10.2	Comparison between our best performance and state-of-the-art segmentation methods.	142
11.1	Summarized Statistics for different estimations compared to ground truth	156

LIST OF FIGURES

Figure	Page	
1.1	The de-identified data retrieved from clinical scans under IRB approval exhibited large variations in the field of view due to types of scanner, protocols of study, or anatomy variance in patients.	3
1.2	Examples of kidney in different contrast phase CT scans. (a) The pre-contrast CT does not show the substructure of kidney. In (b) and (c), visible cortex, medulla boundary, as well as the pelvicalyceal system, (c-1) shows a clearer tissue contrast (green box). (c-2) shows the segmentation on three kidney substructures. (d) and (e) are portal venous and delayed phase CT scans show the enhanced ureter.	4
1.3	Axial slice visualization of abdominal CT multi-organ segmentation. The algorithm segments the abdominal slice into liver, spleen, pancreas, gallbladder, kidneys, stomach, portal veins, aorta, IVC, adrenal glands, etc.	6
1.4	Example of CT images (subcutaneous fat in navy, visceral fat in brown).	9
1.5	Two connected transformer blocks, in which contains the normalization, multi-head self-attention layers (W-MSA), multi-layer perceptron (MLP).	12
1.6	Representative images are predicted to associate with comorbidities and ICD-10 codes (phenotype components) identified in each risk category. The red outlines show the pancreas tissue can be different under phenotyping contexts. (1) is from a nominally healthy pancreas group with potential lung infections; (2) is from type I diabetes and other chronic kidney disease patients with atrophic pancreas; (3) is from other metabolic syndromes including type II diabetes; (4) is from patients with weight loss and pancreatitis.	13
2.1	Slice disorder problem in three regions with the unsupervised regression network (URN). The left panel indicates the global location scores along slices indices. The body part regression values (blue dots) are inconsistent in the right panel compared with an ideal linear relationship (red line).	24

2.2	Proposed deep blind unsupervised-supervision network (BUSN). Panel (A) is the unsupervised network with robust regression refinement. Panel (B) is the deep supervised network using the refined prediction scores. After training, only the right panel (B) is required to perform body part regression.	25
2.3	The three rows show slices in chest, abdomen and pelvis regions in the same subject, under same regression score (-3, 2 and 10) with four columns (URN, BUSN, BUSN with Neighbor Message Passing and ground truth). The slice predicted by BUSN with NMP is closer to the reference slice.	29
2.4	A representative subject was evaluated with URN (left) and BUSN (right). Green scatters are inliers of influence to the regression, yellow scatters are outliers of no influence to the distributed data. Darker blue line indicates the normal linear regression on scatters points, lighter blue line is the RANSAC regressor result according to inliers. Left panel presents the single URN regression with amounts of outliers result in failure of linearity nature in chest and pelvis regions. Right panel shows the testing result of BUSN method, the distributed scores follows good linearity in chest, abdomen and pelvis regions in CT scan. In summary, BUSN takes advantage of self-supervised network, which presents better continuity in regression result among neighbor slices and shows scatter plots without number of outliers.	29
2.5	Organ navigation task and organ-wise body part regression analysis: Density maps represents the distribution of each organ in whole-body CT scan. The red box range represent the URN method, while the blue box is the BUSN-plain method, and the green box shows the result in BUSN with neighbor analysis. “*” indicates statistically significant (p-value < 0.01 from paired t-test).	30
3.1	Representative physical variance in annotated anatomies’ landmarks. The right subject demonstrates shorter distance between target landmarks.	34

3.2	The demonstration of multiple contrast enhancement phases data in CT. The arterial phases highlight the aorta, portal venous phase data reaches an optimal brightness of abdomen organs such as liver and spleen, the delayed phase is better at capturing ureter system.	36
3.3	The overview of the framework. The volume is sample to several slices, followed by data augmentation for semi-supervised loss. The images are then encoded by patch and position embedding layers.	37
3.4	Qualitative visualization show the effectiveness of maintaining cross subject consistency of slice scores. A neck, aorta arch, pancreas, and furmer end slice for inter-subjects at same predicted scores are shown.	40
4.1	Schematic illustration on enhanced CT scan classification. Physician (box 1) assigns contrast materials (box 2) to patients before screening. Different contrast phases scans (1) – (5) are acquired by CT scanner (box 3) depend on time periods (middle axes): 1) non-contrast, 0s or without contrast medium, the box shows the consistent contrast among tissues, 2) early arterial, 15-20s after injection, the optimal illuminated aorta is shown in the box, 3) late arterial, 40-50s after injection, the yellow box shows the light aorta while portal veins starts to be bright, 4) portal venous, 70-80s after injection kidney cortex starts to be bright, and 5) Delayed, 6-10 minutes after injection, the ureter is very bright shown in the box. Last, different contrast unknown scans are fed to our network system (box 4), which outputs a phase label (box 5).	45
4.2	The training pipeline of the proposed method. The encoded data in latent space are acquired from the encoder part of generator. The phase identities are concatenated with the intermediate representation before being fed into the decoder part of generator. On the right, the proposed classifier uses both adversarial loss and contrast classification loss on real and synthetic image in the training. . . .	46
4.3	Scan-level confusion matrix result on withheld 20 paired subjects of VGG backbone networks. ResNet-50 (right) achieves higher accuracy than VGG-16 (left) by using stacked convolutional layers and skip connection design.	52

4.4	Quantitative result on withheld 20 paired subject of GAN architectures and 3DSE. CD-GAN achieves consistent superior performance than StarGAN and 3DSE. Order of labels are the same as Figure. 4.3, true labels top-down, predicted labels left-right are noncontrast, early arterial, late arterial, portal venous and delayed respectively.	53
4.5	Qualitative result of the classification. Four samples are randomly selected from each enhanced phase. Green boxes show critical criteria for identifying enhancement types. Estimated class probabilities for the true label are shown.	54
4.6	Representative error prediction in each type of contrast phase. Red boxes show the misleading criteria for mislabels. Predicted class probabilities are shown on both ground truth and mislabels (red) with all methods.	55
5.1	Method framework. Given a CT scan with at high resolution of $0.8 \times 0.8 \times 2\text{mm}$, a low-res section (left panel) is trained with multi-channel segmentation. The low-res part contains down-sampling and normalization in order to preserve the complete spatial information. After the coarse segmentation are acquired from low-res UNet, we interpolate the mask to match the image's original resolution. Next, random patch sampling (mid panel) is employed to collect patches, and patches are concatenated with corresponding coarse segmentation masks. Finally, we trained a patch-based high-res (right panel) segmentation model, the high-dimensional probability maps are acquired from integration of all patches on field of views. Majority vote is used to merge estimates into a final segmentation.	63

5.2	Representative random patches for 12 abdominal organs of a single subject. The patch size is 128x128x48 and 8 samples are shown for each anatomy. Patch size defines the volume of field of view corresponding to organs. Large organs like spleen, liver and stomach cannot be covered until a number of patches are sampled, the patch of 128x128x48 covers most regions of mid-sized organ (kidney, pancreas, and portal & splenic vein), while small anatomies (adrenal glands, gallbladder, and vessels) can be covered by single patch with above size. The patch size effect is explored in an ablation study.	66
5.3	Quantitative results from the testing cohort: spleen to liver (50 patches used). We compare our random patch network fusion method with three baseline approaches (high-res, low-res and hierarchical framework). The high-res method presents result with large variance and outliers in boxplot due to limited field of view in each patch. The low-resolution segmentation performs better than high-res method in mean DSC, which indicates complete spatial information is essential in abdominal organ segmentation. The hierarchical approach increases training resolution in the second step and achieved higher DSC. Hierarchical method’s performance is limited when bounding box is inaccurate from previous levels. Our method achieves overall highest result compared to hierarchical method with significant improvement, “*” indicates statistically significant ($p < 0.01$ from paired t-test). The random patch fusion framework employs advantages from both low-res and high-res settings, and it achieves segmentation without resample postprocessing. In boxplot, small anatomies (gallbladder, esophagus) present higher improvements than large organs (spleen, kidneys and liver), which presents higher median DSC, smaller variance and fewer outliers.	74
5.4	Quantitative result for the testing cohort: stomach to adrenal glands (50 patches used). “*” indicates our method outperforms hierarchical method by statistically significant improvement ($p < 0.01$ from paired t-test).	74

5.5	Same subject qualitative result of our method compared to baseline approaches (spleen to liver). Second and third row presents direct high-resolution and low-resolution segmentation, mis-predictions are shown due to limited field of view, and resampling respectively. The hierarchical method presents smoother boundaries but suffers from truncation due inaccurate bounding box from first step. Our random patch fusion method presents complete segmentation masks with smoother boundaries among structures.	75
5.6	The same subject qualitative result of our method compared to baseline approaches from stomach to adrenal glands.	76
5.7	Boxplot and uncertainty curves on patch strategies. The boxplots on the left presents the DSC coefficients on testing scans of baseline hierarchy method compared to the complete tiling. Complete tiling shows less variance than baseline hierarchy method. Red diamonds present outliers in complete tiling, and baseline hierarchy shows better DSC (green triangles) than complete tiling. Uncertainty plots show means/standard deviations comparison of structural tiling, structural tiling plus random shift and only random shift methods along with averaged patches per voxel. This presents DSC of each experiment with averaged covered voxels from 2 to 50.	78
5.8	Boxplot on three different patch size along x-y axes. The ablation study conducted on three abdominal organs (spleen, liver and pancreas). Patch size range from small (64x64x48), medium (96x96x48) and large (128x128x48). The boxplots show that larger patch sizes perform better than smaller patch sizes. . . .	79
5.9	Boxplot on three different patch size along the z-axis. The experiments are conducted on spleen, liver and pancreas with patch size ranging from small (128x128x36), medium (128x128x48) and maximum (128x128x64). The boxplots also present that larger number of slices perform better than less in the volume.	80

5.10	Qualitative result of three representative subjects. From low to high, we show the segmentation result evaluated by our method. The testing performance on external datasets (top: HEM1538-splenomegaly, bottom: ImageVU-pancreas outliers).	82
6.1	Phenotype embedded segmentation architecture in the training phase. The left diagram shows the embedding network combining image features, predictive phenotyping, pre-existing risk conditions lying in the latent space to be fed into the segmentation model. The predictor is trained for predicting phenotype-dependent feature maps and selecting “similar” cluster assignment, where the phenotype information is not required as input in the testing phase. Right top: encoder for processing phenotype covariates, right bottom: the predictor follows self-training scheme from image feature. Here, SA denotes soft assignment for risk embeddings, R for ReLU, FC for fully connected layers.	89
6.2	Testing performance on ImageVU dataset. Left: Distribution (median and quartiles) of DSC, the predictive phenotyping shows smaller variance and reduces the number of outliers (DSC<0.4). Right: The DSC (mean) comparison with varying K . The performance shows higher improvement as K increase from 1 to 4, then becomes marginal after $K = 4$. * denotes statistically significant under Wilcoxon signed-rank test ($p<0.05$).	92
6.3	Two representative cases. The top subject has potential lung infections and relative normal pancreas tissue. The bottom case has type 1 diabetes with degraded pancreas tissue. The accuracy gain of the diabetes case is larger than the normal case, showing the method’s ability for identifying degraded pancreas tissue. . .	94
7.1	Overview of UNETR. Our proposed model consists of a transformer encoder that directly utilizes 3D patches and is connected to a CNN-based decoder via skip connection.	97

7.2	Overview of UNETR architecture. A 3D input volume (e.g. $C = 4$ channels for MRI images), is divided into a sequence of uniform non-overlapping patches and projected into an embedding space using a linear layer. The sequence is added with a position embedding and used as an input to a transformer model. The encoded representations of different layers in the transformer are extracted and merged with a decoder via skip connections to predict the final segmentation. Output sizes are given for patch resolution $P = 16$ and embedding size $K = 768$.	97
7.3	Qualitative comparison of different baselines in BTCV cross-validation. The first row shows a complete representative CT slice. We exhibit four zoomed-in subjects (row 2 to 5), where our method shows visual improvement on segmentation of kidney and spleen (row 2), pancreas and adrenal gland (row 3), gallbladder (row 4) and portal vein (row 5). The subject-wise average Dice score is shown on each sample.	98
8.1	Overview of our proposed pre-training framework. Input CT images are randomly cropped into sub-volumes and augmented with random inner cutout and rotation, then fed to the Swin UNETR encoder as input. We use masked volume inpainting, contrastive learning and rotation prediction as proxy tasks for learning contextual representations of input images.	102
8.2	Overview of the Swin UNETR architecture.	104
8.3	Shifted windowing mechanism for efficient self-attention computation of 3D tokens with $8 \times 8 \times 8$ tokens and $4 \times 4 \times 4$ window size.	108
8.4	Qualitative visualizations of the proposed Swin UNETR and baseline methods. Three representative subjects are demonstrated. Regions of evident improvements are enlarged to show better details of pancreas (blue), portal vein (light green), and adrenal gland (red).	110
8.5	Qualitative results of representative MSD CT tasks. Average Dice values are illustrated on top of each image. Our model demonstrates more accurate performance in comparison to DiNTS for both organ and tumor segmentation across different tasks.	114

8.6	The indication of Dice gap between using pre-training (Green) and scratch model (Blue) on MSD CT tasks validation set.	115
8.7	Data-efficient performance on BTCV test dataset. Significance under Wilcoxon Signed Rank test, * : $p < 0.001$	116
8.8	Pre-trained weights using 100, 3000 and 5000 scans are compared for fine-tuning on the BTCV dataset for each organ.	117
8.9	119
8.10	Qualitative visualizations of the proposed Swin UNETR and DiNTS on MSD Tasks	120
9.1	Left: visual and 3D illustration of the kidney components. Right: Demonstration of the hierarchical transformer design, the 3D block aggregation is conducted every two hierarchies, blocks at a factor of 8 are merged to perform communication of sequence representations.	126
9.2	Overview of the proposed UNesT with the hierarchical transformer encoder. Block aggregation and image feature down-sampling are performed between hierarchies.	128
9.3	Qualitative comparisons of representative renal sub-structures segmentation on two right (top) and two left (bottom) kidneys. The average DSC is marked on each image. UNesT shows distinct improvement on the medulla (red) and pelvicalyceal system (green) against baselines.	132
9.4	Left: DSC comparison on the test set at different percentages of training samples. Middle: Comparison of the convergence rate for the proposed method with and without hierarchical modules, validation DSC along training iterations are demonstrated. Right: Results on the KiTS19 dataset show the generalizability of the proposed UNesT.	133

9.5	The Bland-Atman plots compare the medulla volume agreement of inter-rater and auto-manual assessment. We show the cross-validation on interpreter 1, interpreter 2 manual segmentation on the same test set. Interpreters present independent observation without communication. The auto-manual assessment shows the agreement between UNesT and interpreter 1 annotations.	134
10.1	Illustration of defining atlas standardize reference for multi-contrast phase CT. The color grid in the three-dimensional atlas space represents the defined spatial reference for the abdominal volume of interest and localize abdominal organs with each contrast phase characteristics. Blue arrows represent the bi-directional transformation across the atlas target defined spatial reference and the original source image space.	137
10.2	Framework of study design. From D0-A to D0-C, the flowchart shows the criteria from initial query to the CT scans used in this study. Dataset 1 was used for training and validating base system, and dataset 3 is held for external testing. Dataset 1 and dataset 2 were used for training and validation the proposed system.	139
10.3	The qualitative representation of the single subject registrations, average mapping and variance mapping of each contrast phases are demonstrated. The contrastive and morphological characteristics of kidney organs are demonstrated in the single subject registration and average mapping of each phases. Small variations are shown surrounding the kidney organs region in the variance mapping, while great variations are located in the diaphragm region nearby with liver and spleen (Color bar is provided in the supplementary file).	143
10.4	The surface rendering of the registered kidney with significant morphological variation are also illustrated. The 2D checkerboard pattern demonstrate the correspondence of the deformation from atlas space to the moving image space. A stable deformation across the change in volumetric morphology of kidney (100 cc to 308 cc) are demonstrated with the deformed checkerboard.	144

10.5	Examples of representative subjects' visualization. Sections surrounded by blue labels are cortex segmentation, the red is the medulla label, and green is the collecting system. The automatic segmentation are acquired from using the final system model, the comparison between automatic result with manual annotations shows that our method achieves comparable segmentation performance that can be used for measurements.	145
10.6	Average template for pancreas atlas. The segmentation atlas in red color is shown in three planner view.	146
10.7	The surface rendering of the registered pancreas with variations. The 3D checker-board pattern demonstrates the correspondence of the deformation from atlas space to the moving image space.	146
11.1	Pipeline for the proposed method and unidimensional linear regression estimation methods. The computer-assisted method in estimation 2 includes two CNN models with a coarse-to-fine framework. Estimation 3 and 4 use measurements of length and width (splenic index) from the ground truth for cc (cubic centimeter) volume estimation.	151
11.2	Demonstration of the measurements from pipelines for estimating spleen volumes. The manual and computer-assisted methods evaluate the spleen volume (estimation 1 and 2). The linear estimates (3 and 4) manually extract splenic diameters along different axes (length and width) from an unlabeled CT scan.	152
11.3	Quality assurance of the deep learning method in estimation 2 with computed tomography. Top row: three representative subjects' slice above state-of-the-art. Middle row: three representative cases with successful segmentation. Bottom row: failure cases where manual correction was required.	153
11.4	Bland-Altman Plot for computer-assisted method (estimation 2), linear estimate with length and splenic index (estimation 3 and 4). On each plot, the x-axis indicates the mean volume between the ground truth and the estimation from computer-aided method. The y-axis shows the difference in volume. 1.96 standard deviation in shown as the confidence interval.	155

11.5	The repeatability and reproducibility between different imaging analyst readers on 40 respective studies. Bland-Altman Plot between estimations (2-4). The mean in difference and a confidence interval of 1.96 standard deviation are shown.	156
A.1	Copyright from arXiv	185
A.2	Copyright from Elsevier	186
A.3	Copyright from LNCS	187
A.4	Copyright from SPIE	188
A.5	Copyright from IEEE	189
A.6	Copyright from Medical Physics	189

CHAPTER 1

Introduction

1.1 Overview

Quantitative measurement of abdominal organs provides clinical indications in diagnosis or assessing treatment responses. For instance, accurately measuring spleen volume can be crucial for myeloproliferative neoplasms (MPN) [3] and splenic volume or splenic index reduction (SVR) is associated with improved overall survival [4, 5]. These sensitivities to structural changes can be clinically framed using computed tomography (CT) or magnetic resonance imaging (MRI). The primary application of CT has been suggested for early detection and characterization of patient abdominal anatomies [6, 7]. CT produces signals which are processed by the scanner to generate cross-sectional images (slices) of the body [8]. This has opened new investigations into abdominal imaging biomarkers for diseases such as diabetes, hypertension, kidney malfunction, and other metabolic syndromes [9, 10, 11, 12, 13].

CT scans are used as a screening tool within various regions of the human body, and the Hounsfield unit (HU), the standardized transformation of the original attenuation coefficient measurement that represents radiodensity at standard pressure and temperature [14], provides quantitative scales for detecting possible tissues. By exploiting the imaging techniques, statistical analysis of abdominal organs can provide useful information in the clinical workflows [15, 16]. However, statistical evaluation of CT by clinicians and radiologists is commonly hindered by limited reproducibility or repeatability [5, 17, 18]. Disagreements in CT measurements can result from various scope of possibilities, such as the sensitivity of manual assessments, imaging protocols, or software differences, as well as the inherent heterogeneity of anatomies.

Prior empirically derived expertise has made substantial strides for defining gold standards of the medical image analysis [19, 20]. Modern statistical approaches, such as machine learning, can increase reproducibility and explore biomarkers for substantial clinical problems by exploiting the learned distribution from those gold-standard labels [21, 22]. These methods regularized energy functions that push the manual ground truth towards data-driven models. More recent machine

learning techniques, including deep convolutional neural networks [23] have been studied to extract deep implicit features. Such models [24, 25, 26, 27] are generally successful in processing large scale of natural images [28] and reducing variability and bias in medical image analysis [29].

The focus of this thesis is on quantitative abdominal analysis with data-centric deep learning and clinical contexts. The core aspects which we propose are how characterizing medical images can explore biomarker discoveries and, on the contrary, how clinical context can improve medical image analysis. Specifically, the development of robust medical image segmentation technique, the study of the variability of medical images, the use of clinical context, and spatial long-range dependencies.

The first section in this thesis covers the introduction of abdominal imaging and the challenges of medical image analysis. The next section introduces quantitatively measurements with deep learning tools such as abdomen region of interests (ROI). The third section studies the variability of CT contrast enhancement phases. The fourth section introduces the deep learning-based medical segmentation method for high-resolution 3D images. The fifth section explores the body atlas for spatial variability and organ characterization. The sixth section illustrates the clinical context of imaging and non-imaging features. The last section introduces transformer models for modeling long-range dependencies of 3D medical images. Last in this thesis, the prior contributions, and recent challenges regarding the field are outlined.

1.2 Abdominal Imaging

Computed tomography is routinely used for abdominal imaging in the clinical practice [30]. CT images use fan-beam X-rays for demonstrating various anatomic structures with high-contrast based on physical densities, resulting in absorbing/blocking X-rays and intensity of CT images (i.e., Hounsfield Units), in which the radiodensity of distilled water at standard pressure and temperature (STP) is defined as, while the radiodensity of air at STP is defined as -1000 HU [14]. The standard attenuation have a relatively consistent intensity range for anatomical tissues (e.g., bone appears brighter than soft tissues), which allows clinicians and patients to see the interior body without surgery. Despite the consistent scale of HU, this section demonstrates the variability of CT imaging from perspectives and the challenges for conducting robust automatic methods of abdominal tissues on a large scale of clinical data.

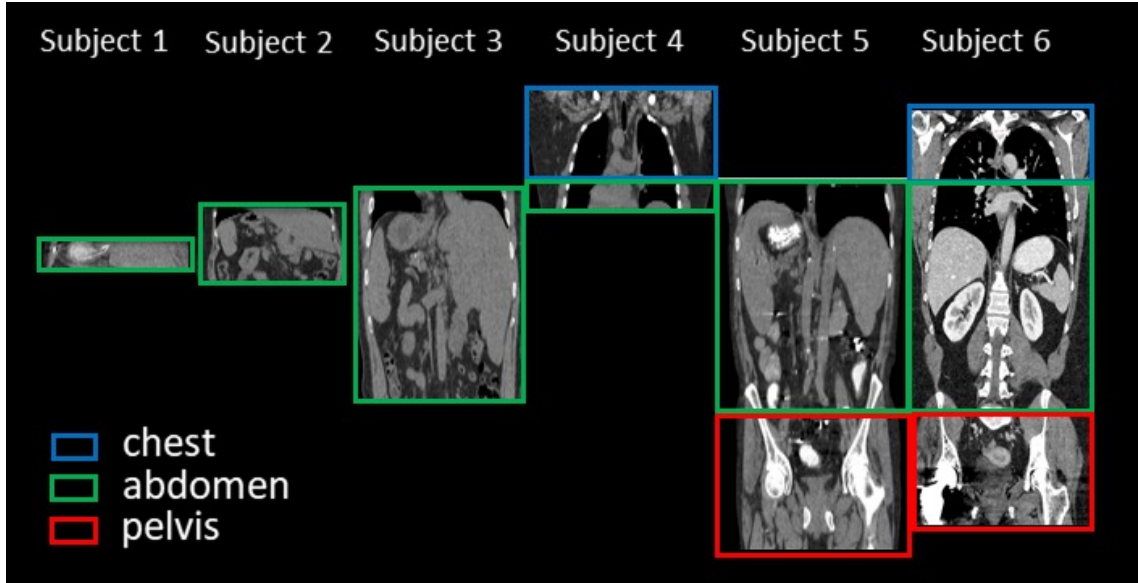


Figure 1.1: The de-identified data retrieved from clinical scans under IRB approval exhibited large variations in the field of view due to types of scanner, protocols of study, or anatomy variance in patients.

1.2.1 Abdomen CT Field of Views (FOVs)

Clinical CT scans are acquired protocols. Different requirements exhibit large variabilities in image resolution, contrast, sequence, and patient anatomies. Field of view (FOV) is a critical property for volumetric images, along the cranial-caudal axis, three-dimensional data can have a range of pelvis or thigh to lung areas. Regarding anatomies, abdominal CT scans include from the upper esophagus, aorta through liver, spleen, pancreas to the lower kidneys. Referencing body parts, abdominal imaging scans body waist, abdominal viscera, outer abdominal wall, muscle, inner abdominal wall, subcutaneous fat, organs, and other soft tissues. The diversity of coverage raises challenges of structural measurements due to spatial alignment between subjects as shown in Figure. 1.1.

To quantify body consistency from variations of FOV, especially patients with metabolic syndromes, prior study [31] suggested anatomical structure localization (i.e., axial slices) that can be identified by spatial consistency. The potential applications include content navigation [32], lesion detection [33], classification [34], and segmentation [25, 35], which universally benefit from accurate quantitative assessment of body parts in regions of 1) shoulder and lung, 2) abdomen and 3) pelvis.

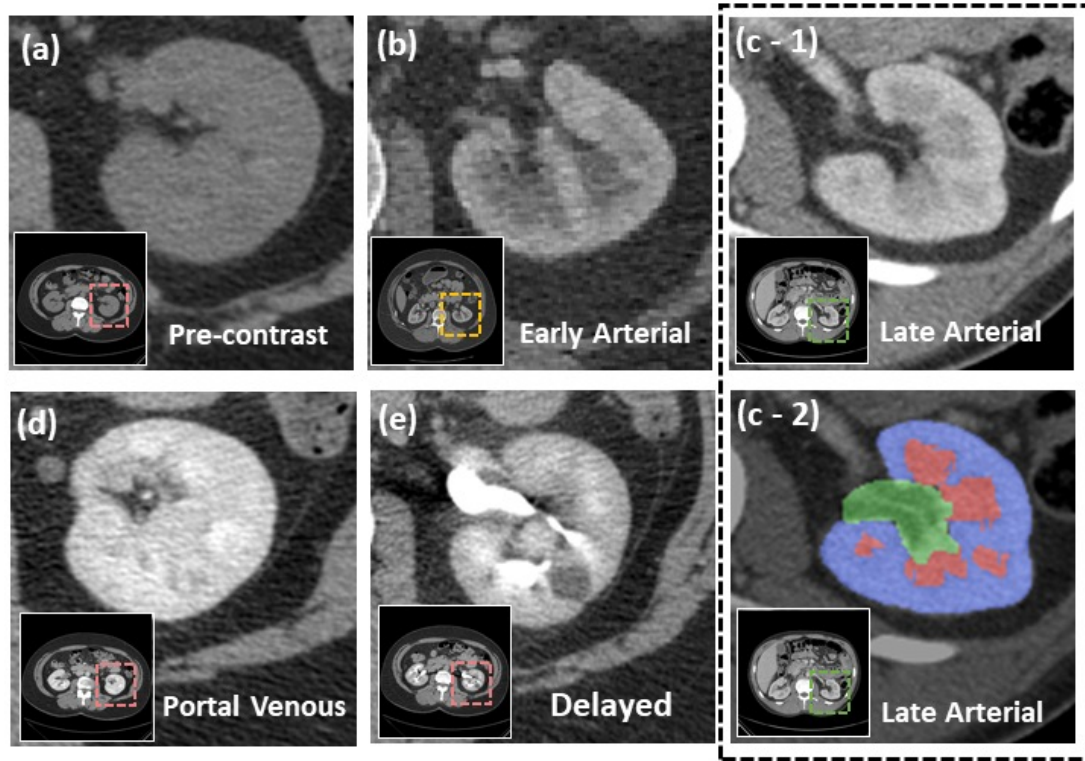


Figure 1.2: Examples of kidney in different contrast phase CT scans. (a) The pre-contrast CT does not show the substructure of kidney. In (b) and (c), visible cortex, medulla boundary, as well as the pelvicalyceal system, (c-1) shows a clearer tissue contrast (green box). (c-2) shows the segmentation on three kidney substructures. (d) and (e) are portal venous and delayed phase CT scans show the enhanced ureter.

1.2.2 CT Enhancement Phases

Dynamic CT images are widely used for screening abdominal organs [8]. Radiocontrast agents are suggested [36] in the clinical process to improve the visibility of internal anatomies. Contrast medium (i.e., intravenous iodine) is recommended by physicians before scanning [37]. During scanning, maximal contrast differences can be acquired according to the timing. In addition to non-contrast CT were typically used for detecting calcification, fat in tumors or inflation, enhanced phases [38, 39, 40, 41] typically contain 1) early arterial phase CT, where the scan is obtained after bolus tracking for optimal contrast of arteries, 2) late arterial phase CT, in which enhancement of the portal vein can be seen, 3) portal venous phase CT, acquired about 1 minute after injection, can provide optimal contrast of liver, spleen through blood supply by the portal and splenic vein, 4) delayed phase CT, which is scanned 6-10 minutes after bolus tracking, show the best contrast

difference of renal structures such as renal veins and urinary system. Figure. 1.2 shows an example of the renal cortex, medulla, and pelvicalyceal system can be better visualized at late arterial phase CT scans compared to other phases.

1.2.3 CT Resolution

The resolution of clinically acquired CT scans can have potential differences. Among the cohorts that we have been studying with, the slice thickness (along with the cranial-caudal view) approximately ranges from 0.8 mm to 8.0 mm, while the in-plane resolution ranges from 0.5 x 0.5 mm to 1.0 x 1.0 mm. Conventionally, the voxels of the CT volumes are highly anisotropic (1.0 x 1.0 x 3.0 mm), the large slice thickness makes it difficult to yield smooth 3-D surfaces of anatomic structures. In addition to slice thickness, the variability of multiple views' reconstruction also can have a substantially diverse resolution. During the reconstruction process, given considerations of radiation dose, acquiring a volume acquisition CT is the ability to reconstruct the images in three planes: the axial, sagittal, and coronal planes. Viewing the abdominal organs in all tri-planes is particularly useful. Typically, the axial reconstructed scans are commonly used for image processing and analyses due to the consistent anisotropic planes with 512×512 pixels. In real-world medical image analysis applications, interpolation methods (e.g., bilinear interpolation, nearest interpolation) are implemented for scans to keep consistent resolution and dimensions. Under current approaches [42, 43, 44], high resolution (approximately $0.8 \times 0.8 \times 1.0$ mm) is used for quantitatively evaluating abdominal organs.

1.3 Body Compositions

1.3.1 Abdominal Organs

As below, 13 main abdominal organs are demonstrated, and viscera backgrounds inside the abdominal cavity are also shown.

Spleen: In the upper left abdomen, the spleen is a large organ working as a blood filter in the human body. The spleen removes old red blood cells and keeps the reserve of blood. In enhanced CT images, the spleen exhibits a slightly brighter contrast in the venous phase. Neighboring tissues are the kidney, splenic artery and veins diverge at the right side of the spleen. Spleen diseases are associated with many syndromes [45, 46, 47] and can be an essential marker for liver, pancreas

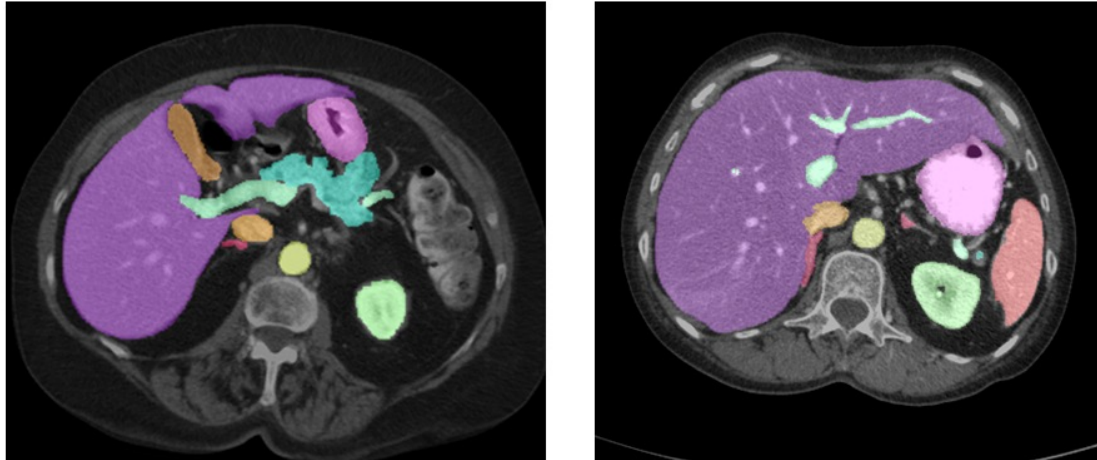


Figure 1.3: Axial slice visualization of abdominal CT multi-organ segmentation. The algorithm segments the abdominal slice into liver, spleen, pancreas, gallbladder, kidneys, stomach, portal veins, aorta, IVC, adrenal glands, etc.

disease, cancer, and other infections [48].

Kidney: Kidneys work as essential organs in urinary systems, they serve as filters of the blood by removing wastes to the urinary bladder. The kidneys locate in the lower abdomen, strictly, they are posterior to the rear of the abdominal cavity in the retroperitoneum. Kidneys are typically connected to the rear of the liver and spleen, around most of the air HU (≈ 1000), showing a relatively clear boundary. Multiple contrast protocols are suggested for screening kidneys, specifically non-contrast, portal venous, delayed phased are used. Inside kidneys, renal structures can suggest as useful markers for evaluating kidney morphologies. Including renal cortex, columns, medulla, and pelvicalyceal system, measurements of these inner anatomies can be opportunities and challenges for investigating renal functions.

Gallbladder: The gallbladder is a glandular organ where bile is stored before being released into the small intestine. It locates beneath the middle liver; a normal gallbladder can have liquid contrast showing lower contrast to the liver. Some patients have gallstones that appear brighter than surrounding tissues.

Esophagus: The esophagus is the organ that contains the muscular tube through digesting systems from the pharynx to the stomach. In the study, the esophagus is the up-most organ that ranges from the middle thorax region to the middle abdomen.

Liver: The liver provides a wide range of functions, including protein synthesis, production of biochemicals, and detoxification. It is located under the diaphragm (upper right quadrant of the abdominal cavity). The inferior vena cava passes through the right posterior of the upper liver; portal vein drains into the liver from middle abdomen; gallbladder underneath and connects to liver.

Stomach: The muscular, hollow organ performs the digestive function of the human body. The shape of it depends on the patient's food consumption before scanning. In addition, the contrast of stomach is also potentially depending on patients. It locates between the esophagus and the small intestine, the left upper region of the abdominal cavity.

Aorta: The aorta is the essential vascular vein that distributes oxygenated blood to the human body. Originating from the left ventricle of the heart to the lower abdomen, then divided into two smaller arteries. Aorta can be observed in normal enhanced CT images and brighter in arterial phase scans.

Inferior Vena Cava (IVC): The IVC is located in the posterior of the abdominal cavity and on the right side of the vertebral column. It distributes de-oxygenated blood from the lower body to the right atrium of the heart. Parts of IVC pass through the liver and appear brighter in contrast to surrounding tissues in enhancement phases.

Portal and Splenic Vein: The hepatic portal vein carries blood from the gastrointestinal tract and spleen to the liver. The portal and splenic veins are formed by the connection of the superior mesenteric vein and vessels in the spleen. The middle portal vein is substantially surrounded by the pancreas, and the portal vein connects the liver, splenic vein connects the spleen. In contrast-enhanced CT scans, portal and splenic veins appear brighter than the liver and spleen, especially in the portal venous phase.

Pancreas: The glandular organ is responsible for many essential functions, including producing hormones, secreting pancreatic enzymes for digestion. The pancreas is composed of the head, body, and tail, each part's function can be different. The head is typically located in the concavity of the duodenum, the body behind the stomach, and horizontally extends along the splenic vein, the tail directs to the spleen. The pancreas is one of the most morphological organs in the human abdomen, the shape can be variant depending on patient diseases.

Adrenal Glands: The left and right adrenal glands are endocrine glands responsible for releasing hormones. Adrenal glands are relatively small anatomies in the abdomen cavity, they are

wishbone-shaped structures. The intensity and texture are similar to the glandular organ – the pancreas. They are located in the upper kidneys, the right adrenal gland is close to the lower-left liver, and the left is close to the lower right spleen. Segmentation examples are shown in Figure. 1.3.

1.3.2 Body parts

Other anatomic structures are important for quantitatively measuring abdominal contexts. Understanding comprehensive backgrounds can be supportive to the clinical analysis, especially for abdominal metabolic syndromes. Meaningful backgrounds include:

Body mask: The body mask covers all areas inside the human body, differencing air intensities in the CT scan. **Fat:** The skin surface or the body waist circumstances defines the outline of the human body. Outside the abdominal wall, and muscle are the subcutaneous fat. Inside the abdominal wall, visceral fat is surrounded by most soft tissues including organs. Subcutaneous fat and visceral fat are crucial for measuring the health status of patients, obesity levels.

Muscle: The oblique abdominal muscles are composed of three types: the external oblique, internal oblique, and transverse abdominal. Muscles form the lateral boundaries of the abdominal wall and are brighter than the surrounding fats and inner tissues. The surrounding fats such as subcutaneous fat are outside abdomen cavity, the visceral fat around abdominal organs are in the abdomen wall.

Abdominal wall: The abdominal wall is an essential structure to characterize ventral hernias, and to differentiate muscle, bones. The abdominal wall defines the abdomen cavity, most organs locates inside the inner abdominal wall. Examples of body compositions are shown in Figure. 1.4.

1.4 Large Scale Medical Data

The large-scale data resources of medical images and electronic health records (EHR) can be collected within institutes. For leveraging innovative programs for clinical and translational research, institutes established platforms for enhancing the quality and efficiency of research conducted. Such as ImageVU, the research data archiving platform, which serves as Vanderbilt's reserve of computed tomography (CT) and magnetic resonance imaging (MRI), along with image metadata including acquisition protocols, study descriptions, and exam codes.

Variations in Large-Scale Data: As high-performance computers are widely used, the standard

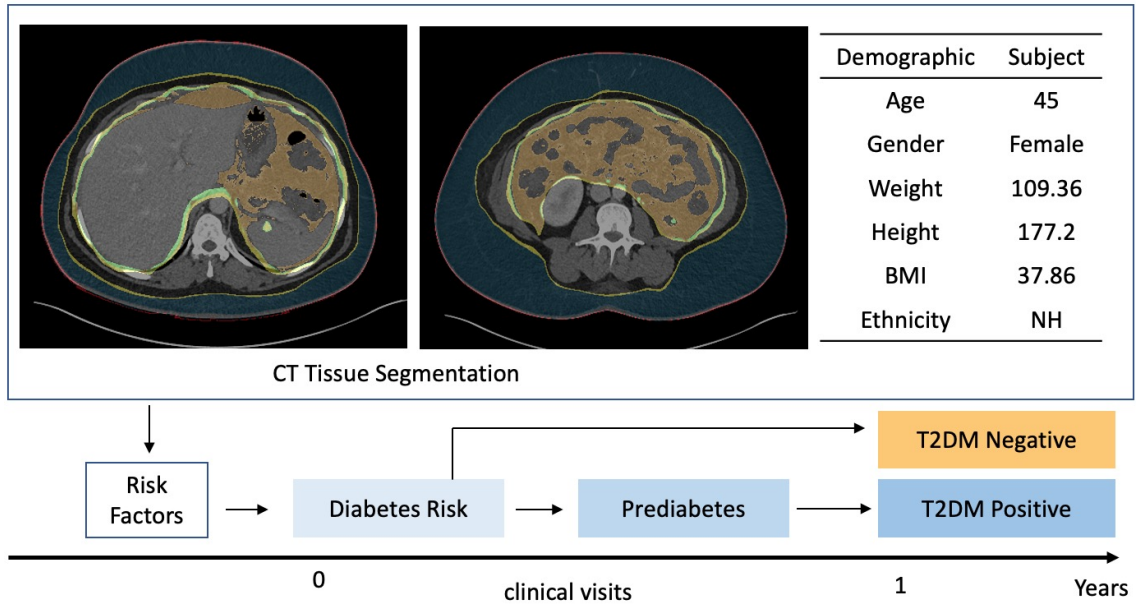


Figure 1.4: Example of CT images (subcutaneous fat in navy, visceral fat in brown).

processing for medical images along with the powerful data platforms are investigated to enable high throughput analysis. In addition to the individual scale of the study, the works also provide visions in large-scale populations to understand human body functions, behavior, disease, and syndromes. The large-scale data analysis brings new challenges as patients are heterogeneous, variations pose difficulties in rectifying inter-and intra-subjects. Large statistical models are proposed to reconcile such variation; however, there are other challenges such as variations in scanning protocols, modalities, tissue morphology, etc, which are not well considered in related works. Considering intra- and inter-subjects, several essential problems still exist, such as image contrast/intensity standardization, region of interest's navigation, quality assurance, denoising, and robust registration.

1.5 Machine Learning in Medical Image Analysis

To quantitatively and qualitatively investigate medical data, modern machine learning concepts and techniques are essential. The section covers major machine learning approaches for solving problems in clinically acquired medical images. In addition, the chapter describes how these techniques are used for volumetric analysis and application to clinical workflows.

1.5.1 Medical Image Segmentation

Segmentation refers to the process of acquiring target objects (i.e., organs) at pixel-level from an image, is one of the most widely used tools for quantitatively and qualitatively assessing organs, tissue in the medical image analysis [49, 50, 51]. Modern approaches are discussed from multi-atlas methods [19, 52, 53, 54, 55, 56], which uses a pairing of scans and corresponding manual labels. The typical framework of atlas-based methods is: 1) a set of annotated labels are registered to a target subject, and 2) label fusion to construct the moving image segmentation. Another most adopted approach is the deep learning [57, 58, 59, 60, 61, 62, 63, 64], which is described as a supervised learning paradigm using expert annotated labels as ground truth data. The models are trained to recognize patterns and information and minimize the cost function of prediction and ground truth. In medical segmentation, the model predicts to assign a label class to every pixel in a scan. Pioneering works such as UNet [57, 65] and FCN [66] successfully characterized an encoding step and decoding step, where the more extract global features into representation, then upsample low level features back to original image space. Inspired from fundamental computer vision to medical images, medical image segmentation accuracy has achieved significant improvement in recent years even though expert efforts are very time-consuming and resource-intensive.

Due to the limitation of computing hardware, such as memory of GPU, 3D CNN-based methods are rarely explored. This results in impractical applications as the spatial context of medical images is important. Instead of feeding the entire scan into hardware, researchers proposed patch-based algorithms [42, 43, 63, 67] for adapting high-resolution CT/MRI scans.

1.5.2 Self-Supervised Learning

Deep neural networks have been used as the fundamental architecture [24, 68, 69] to many target tasks such as detection, classification, and segmentation. The models trained from expert annotated data like ImageNet show superior results. However, the performance of these models is common concerning the amount of training data, different networks were designed with denser architectures including AlexNet [70], ResNet [24], DenseNet [71]. Collecting large-scale annotated datasets in the medical domain is resource-intensive. The BTCV dataset [72], which consumes more than 10 experts' efforts forms 30 subjects' multi-organ segmentation labels. To exploit the large-scale medical images and avoid extra expert efforts, self-supervised methods were proposed to learn

visual and contextual features from unlabeled images.

In self-supervised learning, the models treat the learning as a supervised learning task with the generated pseudo-label, which is called pretext tasks. By modeling from pretext tasks, networks are trained by optimizing designed objective functions. Recent studies on pretext tasks of natural images have been focused on colorizing [73], inpainting [74], jigsaw puzzle [75], etc. There are two key factors for designing pretext tasks: 1) the augmentation/distortion can be automatically generated without manual efforts; 2) the visual features captured from the images are useful for representation learning.

Self-supervised learning and transfer learning [76] are explored in recent years and becoming the de facto framework [76] for medical image analysis. Prior works [77] focused on chest CT such as LUNA [78] due to its promising application of identifying lung nodules. Other approaches like denoising [79], image inpainting [74], jigsaw puzzle [75], clustering [80], patch shuffling [81], Rubik's cube [82] are proposed, which achieved milestones in downstream tasks. More recently, the contrastive learning [83] brought insights and show its powerful capabilities in capturing visual representations. Contrastive learning is conducted between augmented data pairs, the model attempt to differentiate positive or negative pairs by minimizing contrastive loss.

1.5.3 Transformers

Self-attention-based transformer models [84, 85, 86] are a family of prominent deep learning methods. Initially, self-attention is a mechanism aligning different sequence positions to compute a hidden state, and it has become an integral part of sequence modeling and transduction methods. Due to its advantages on learning dependencies of long sequences, the transformer is used for processing sequential signals such as the natural language processing (NLP) [86] or speech processing [87]. The original receipt of transformer consists of an encoder and a decoder [84], each part has a stack of L blocks of the self-attention module as shown in Figure. 1.5.

Inspired by the major success of NLP, researchers brought the idea to computer vision. The vision transformer (Vit) [88], which partition image into patches and flattened sequences, directly transferred the advantages of the sequence-to-sequence model to image recognition. The exceptional performance attracts works in the object detection field [89], semantic segmentation [90], and video understandings [91].

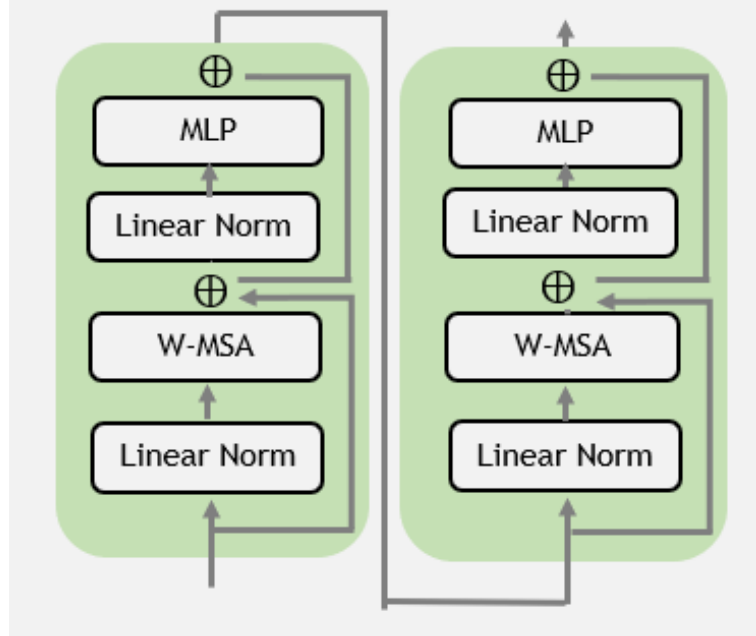


Figure 1.5: Two connected transformer blocks, in which contains the normalization, multi-head self-attention layers (W-MSA), multi-layer perceptron (MLP).

There are two major advantages of transformer-based models, 1) transformers provide the larger capacity for learning long-range dependencies and show their effectiveness of capturing global contextual information; 2) they can benefit from large-scale pre-training.

Transformer models also demonstrate its superior transferability [88] from pretext tasks to downstream tasks. As discussed in the prior section, collecting large-scale annotated data is non-trivial, expensive, and time-consuming. Researchers are exploring better models for the capabilities of capturing features through self-supervised learning. Empirical experiments show that transformer models are superior at modeling large-scale data and preserving its representations into broader target tasks. This advantage attracts researchers in the medical domain as data without expert annotations are easier to get.

The rapid increase of transformer-based models motivated the medical image analysis [92, 93, 94], fundamental medical image problems are under revisiting with self-attention mechanisms in a sequence to sequence manner. In these pioneering works, the transformer blocks are used either as a bottleneck feature encoder or as additional modules after the convolutional layer achieved potential promising results.

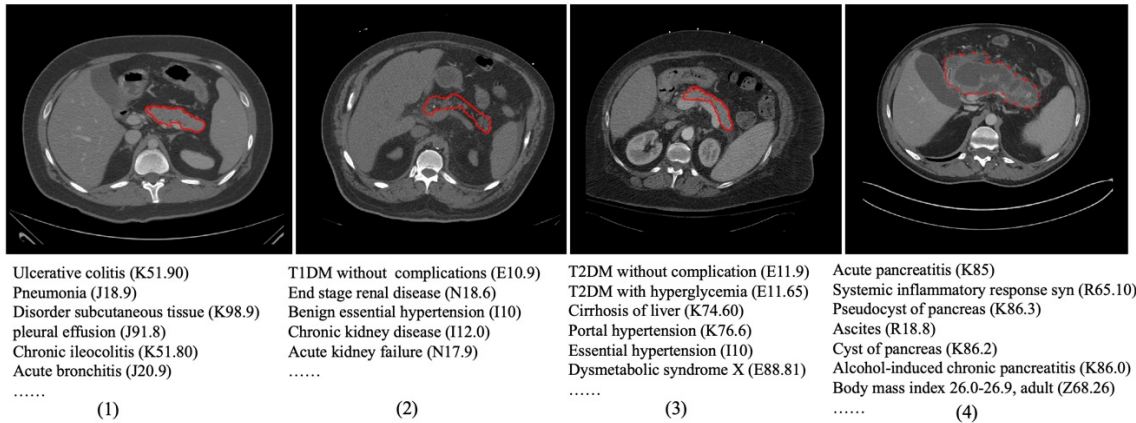


Figure 1.6: Representative images are predicted to associate with comorbidities and ICD-10 codes (phenotype components) identified in each risk category. The red outlines show the pancreas tissue can be different under phenotyping contexts. (1) is from a nominally healthy pancreas group with potential lung infections; (2) is from type I diabetes and other chronic kidney disease patients with atrophic pancreas; (3) is from other metabolic syndromes including type II diabetes; (4) is from patients with weight loss and pancreatitis.

1.6 Multi-Modal Clinical Context

Electronic health records (EHR) [95] associated with patients' healthcare journey including demographics, procedural terminologies, diagnoses, lab tests, and medication notes. These non-imaging clinical contexts are categorized into structured (i.e., ICD, CPT codes) and unstructured formats (i.e., notes). The major usage of EHR is to efficiently query or access healthcare history for clinicians in decision making and diagnostic processes such as Parkinson's disease and type 2 diabetes (T2D) [96, 97]. However, the ability to identify specific disease progression is often limited because the feature extraction approaches require manually labeling, resulting in limiting complexity and biasing to feature significance. Additionally, feature selection from a single modality (e.g., EHR or image) is a highly variable process that relies on tuned parameters. Recently, many applications [98, 99, 100] are found in the biomedical informatics and epidemiology for precision medicine. Patient trajectory modeling, outcomes prediction [101, 102] is studied with the data-driven algorithms shown in Figure. 1.6.

1.6.1 Predictive Modeling with EHR

Predictive modeling with bioinformatics data drives the quality of medicine with personalized and quantified information. Fortunately, the routinely collected patient electronic records are approach-

ing a large scale in volume and complexity. However, these EHR data are yet to be used in constructing predictive models to support care delivery. First, querying and formatting EHR data is difficult since statistical models require customized variables to analyze specific disease progressions. Next, preprocessing, curating variables needs many efforts and expertise, the inappropriately collected variables may result in biased or imprecise predictions. In addition, EHR data are longitudinal in nature, missing visits or medical events can lead to poor clinical outcomes. Furthermore, EHR exhibits features as a single modality, incorporating imaging biomarkers can offer a better understanding of patients.

Recent studies of deep learning unlock mining useful information from EHR [103, 104]. Key deep learning architectures (e.g., convolutional neural networks (CNN) [105], recurrent neural networks (RNN) [106]) have been employed for the modeling of EHR data. From Boltzmann machines [107] that learning a representation in an automatic manner to denoising autoencoders for extracting features, they showed the promising power of reducing the tedious expert efforts. In both methods, logistic regression, random forest, or support vector machines (SVM) are used under supervised learning to map representation to the outcomes. Considering the longitudinal nature of patient data, RNN exhibits early success in modeling temporal information. To predict future disease progression based on EHR or imaging trajectories, long short-term memory (LSTM) [108] with a forgetting mechanism are introduced to handle irregular time intervals of each variable. In this article, we aim to analyze the real-world challenges when incorporating complex EHR. The machine learning models that adapt the data in the wild can show some insights into best practices in clinical prediction tasks 1.6.

1.6.2 Multi-modal Representation Learning for Medical Data

The multimodality learning [109, 110] has been widely studied in recent years because of its ability to utilize different types of input data. Modality specifies a single target task with varied representation formats, and multimodal learning suggests understanding with a variety of senses engaged in the modeling. Inspired by the pioneer works in the NLP, natural image domain, researchers explored the usage of semantic knowledge of medical images along with diagnostic reports for explanatory supports [111, 112].

1.7 Challenges in Modern Medical Image Analysis

The application of medical image analysis to big clinical data is a challenging field with recent technological advances. However, most prior works directly adopted brutal machine learning algorithms, clinical contexts are inadequately involved in the inferencing process. In this thesis, we target the challenging task of exploiting clinical context into data-centric deep learning techniques. Including those published in this dissertation, we aim to facilitate the ability to investigate many yet unanswered questions about the theories, methods, and translational applications. Starting from critical challenges of clinical acquired large-scale data, we summarize key challenges and our aims as below.

1.7.1 Variations in Medical Image FOV

Clinically acquired CT scans can exhibit large variations in the field of view, especially for coverage of the chest, abdomen, or pelvis. Without pre-processing, such scans are difficult to use for medical image analysis due to a lack of spatial consistency. Using pre-processing to remove inconsistency in the field of view helps to localize anatomical regions in each body part and enable more precise image registration and machine learning. Recent method [31] suggested that image quality, quantitative analysis, and anatomical structure localization could be combined to estimate spatial consistency across the human body in CT. The method required the creation of distance metrics, varied combinations of scalars, isotropic variance, and slice thickness, which resulted in difficulties in creating a consistent performance for capturing continuities between slices in medical image volumes.

Aim 1: To achieve slice-wise tissue navigation and to quantify body consistency, we aim to explore the robust body part regression technique with self-supervised learning, which estimated a uniform spatial location. We target the slice-wise regression across the human body for CT and MRI images.

1.7.2 CT Enhancement Phase Identification

Dynamic contrast enhancement Computed Tomography (Dynamic CT) is widely used in clinical diagnosis. For quantitative measurement in tissues, the standard Hounsfield Units (HU) scale is used in contrast-enhanced CT scans for describing radiodensity. However, the differences and metrics

are often investigated manually by physicians. Therefore, missing information or mislabeling are commonly observed in large-scale studies. Revisiting phase knowledge and contrast protocols in meta-data are resource-intensive. Moreover, correcting phase labels is hard and challenging due to variations in tissue, contrast material, injection protocols, vascular dynamics, and metabolism.

Aim 2: To investigate the automated method of identifying CT contrast phases, we aim to propose a multi-domain contrast disentangling GAN, by learning disentangled representations across contrast phases features. The method needs to learn an intermediate representation, then reconstruct a synthetic contrast-enhanced image. The goal of using a generative model is to employ adversarial loss and synthetic classification as data augmentation for improving classification performance

1.7.3 High-Resolution 3D Medical Image Segmentation

Current medical image segmentation methods [26, 64] are inspired by the segmentation method from natural images. They perform well on smaller dimensional 2D images, however, medical images are high resolution and high dimensional or 3D volumes. Deep neural networks can suffer from limits of computing resources. Regarding the resource-accuracy trade-offs, 2D approaches are typically taking separated slices for training resulting in lacking spatial information. The 3D architecture needs scans to be either 1) patched or 2) down-sampled, it is the slowest way to train at approximately 80 iterations per minute and one patch per iteration.

Aim 3: To exploit 3D context and to cope with the limitation of computational resources, researchers investigated hierarchical frameworks, we aim to propose a concise coarse-to-fine framework by using random patch network fusion to alleviate the difficulties for 3D multi-organ segmentation. The method is needed to segment original CT without image scaling in the input and output. We propose a two-stage cascade architecture that utilizes the fact that the performance of a higher resolution level in a hierarchical model is indicative of the low-resolution level in the hierarchy.

1.7.4 Body Reference Space

Studies have shown great impact and efforts on building brain region atlases for clinical usage with different modalities images. However, limited studies are proposed in creating a standard reference framework for abdominal organs and showed that the development of abdominal atlases is remained challenging across multi-modality images (CT and MRI).

Aim 4: We target to present a contrast-preserving CT abdominal atlas framework, optimized for healthy kidney organs with contrast-characterized variability and the generalizable features across a large population of clinical cohorts. Meanwhile, we explore the study in kidney substructures including renal cortex, medulla, and pelvicalyceal system. We propose a pioneering work to segment each part of the kidney substructure and conduct volumetric analysis for reproducibility.

1.7.5 Combining Clinical Context

The ability to phenotype patients is a fundamental task in clinical research and medicine. Current methods rely primarily on bioinformatics data such as EHR markers. In addition, previous approaches include heavy manual efforts of feature engineering. However, considering the heterogeneous condition of patients, the utility of EHR may not be sufficient to provide support of accurately subtyping. Such as type II diabetes progressions are prognosed on the expression of lab tests, demographics (e.g., BMI), and other imaging markers (e.g., a retinal screening, pancreatic abnormality). Therefore, EHR-based approaches may be undesirable because 1) biomarkers lack specificity and comprehensiveness, 2) the sample or patient may be too heterogeneous, 3) several markers required can be larger than available EHR detected. In many of these conditions, a critical challenge is whether multi-modal data could be modeled for comprehensive analysis.

Despite the rapid development in multimodal deep learning research, limited related works are observed in the medical domain. Partly due to the complexity of anatomies and scarcity of well-organized clinical data, the paired imaging features, and informatics features are hard to interact with. Previous approaches explored data-level fusion in a canonical way such as early, late or intermediate fusion. Different strategies are used during the decision-making process because of the varying assumption of conditional independence between multiple data modalities.

Aim 5: To address above challenges, we target the pancreas segmentation task to model both the pancreas imaging features and clinical features via predictive phenotyping. The rationale is that the larger scale of EHR data with (e.g., ICD-10 code) which indicates phenotype subgroups can be potentially correlated to the different appearance of the pancreas. Specifically, the proposed approach consists of an encoder, a segmentation decoder, and a predictor with sets of phenotypes candidates' centroids.

1.7.6 Modeling Medical Spatial Representations

Although such CNN-based approaches have powerful representation learning capabilities, their performance in learning long-range dependencies is limited to their localized receptive fields. As a result, such a deficiency in capturing multi-scale information leads to sub-optimal segmentation of structures with variable shapes and scales (e.g. brain lesions with different sizes). However, the locality of the receptive fields in convolutional layers still limits their learning capabilities to relatively small regions.

Aim 6: To leverage the power of transformers for volumetric medical image segmentation and introduce a novel transformer-based architecture. We aim to reformulate the task of 3D segmentation as a 1D sequence-to-sequence prediction problem and use a transformer as the encoder to learn contextual information from the embedded input patches. The extracted representations from the transformer encoder can be merged with the CNN-based decoder via skip connections at multiple resolutions to predict the segmentation outputs.

1.8 Contributed Work

We have developed tools for quantitative image biomarkers with 3D medical images. This includes construction of CT-based body part regression (contribution 1), contrast identification model (contribution 2), and high-resolution medical image segmentation (contribution 3). In addition, we study the medical image analysis with clinical context, which contains phenotyping with non-imaging markers and outcomes predictions to diabetes patients (Contribution 4). Subsequently, we combine recent advanced self-attention mechanisms in medical image segmentation tasks (Contribution 5) for modeling spatial long-range dependencies with self-supervised learning.(Contribution 6). The 3D atlases is built to support tissue mapping for kidney, kidney substructures, and pancreas (Contribution 7). Finally, the application of the quantitative biomarkers are demonstrated for splenomegaly analysis (Contribution 8).

1.8.1 Contribution 1: Body Part Regression for CT

Computed tomography (CT) is an essential imaging technique for human body scanning. However, there are challenges regarding the quantitative and qualitative analysis. We investigate the computer-aided approaches for modeling a large degree of variability in the abdominal CT images.

We evaluate the reproducibility and sensitivity of the automated methods for organs. Body compositions are varying significantly in medical images which results in a gap between clinical acquired image and research quality image. This is expected because the raw data from the medical center are acquired in a large variety of parameters. The previous method exhibits severe nonlinearities with the order of slices. To curate the data in the wild, we investigate the use of spatially navigation techniques characterized as body part regression and evaluated the approach for abdominal organ analysis.

The contributed aspects are covered in the following, we proposed (1) a self-supervised solution to boost current body part regression are designed using 1030 multi-center CT scans without using manual labels; (2) a novel unsupervised-supervision method is introduced to achieve robust body part regression and (3) a preprocessing pipeline is proposed using BUSN to normalize the CT volumes spatially, which is evaluated by organ navigation and 3D multi-organ segmentation on normalized scans.

This part will be covered in Chapter 2 and Chapter 3.

1.8.2 Contribution 2: CT Contrast Phase Identification

The dynamic CT images are heterogeneous, showing scanner-dependent variation to meet the demand of contrast optimum. However, the difference in scanning protocol, morphologies in organs, and metabolic effect variation in patients are significant. These unknown factors potentially impact the quality of medical image analysis. To tackle the CT contrast characterizations, we propose a bi-level optimization approach to classify the contrast phases in CT scans. In this work, we propose CD-GAN, an adversarial learning network to perform contrast phase classification by learning from both real and synthetic images. 1) A two-step generator is introduced, which learns the disentangled representations and domain-specific features among multiple contrast CT. 2) We perform quantitative and qualitative evaluations on the prevalent baselines (VGG, ResNet-50, StarGAN, and 3DSE).

This part will be covered in Chapter 4.

1.8.3 Contribution 3: High-Resolution 3D Medical Image Segmentation

Medical image segmentation is an essential way to quantitatively access imaging biomarkers. After retrieving the IRB-approved medical images, we propose the computer-aided method for high resolution 3D CT scans. Regarding abdominal organs, body circumstance, fats, muscles, etc, we map a field of view (FOV) into deep learning models outputs. The contributions of this work are: 1) We proposed a new coarse-to-fine framework termed ‘random patch network fusion’ by introducing randomly localized patches between first and second stage. 2) We show that our proposed method can be implemented to predict original space segmentation in second level model. 3) We provided large-scale validations on analyzing patch-based strategies and comparing them with our method, supporting that patch-based method plus random shifting could boost 3D segmentation performance.

This part will be covered in Chapter 5.

1.8.4 Contribution 4: Combining Non-Imaging Clinical Context

Comprehensively interpreting medical data is becoming challenging because of heterogeneity and variability of data modalities. In the past works, we studied outcome predictions with large amounts of variables, such as phenotyping with interpretive patient information. We also investigated disease progression associated with EHR data, which conveys important information. In particular, we proposed a universal framework for pancreas CT segmentation with knowledge by EHR data. Our work empowers machine learning models with interpretive abilities, which to 1) identify patient phenotypes across imaging and EHR data, 2) bridging clinical feature modalities to representation learning.

This part will be covered in Chapter 6.

1.8.5 Contribution 5: Spatial Long-range Dependencies

Transformer-based models show promising representation learning power for NLP and vision tasks. In medical image analysis, there are emerged 2D and 3D approaches that achieves advantages on classification and segmentation tasks. Transformer models also show its superiority of benefiting from pre-trained weights, which can facilitate target tasks by reducing burden of manual annotation. We proposed to use transformer encoder to increase the model’s capability for learning long-range

dependencies and effectively capturing global contextual representations for 3D medical images.

This part will be covered in Chapter 7.

1.8.6 Contribution 6: Efficient 3D Transformer Models with Self-Supervised Learning

Transformer-based models are of better capability benefit from large-scale pre-training. Instead of acquiring massive gold-standard labels, we propose to use unlabeled data with the self-supervised learning technique to model representations using image inpainting, contrastive learning and rotation prediction. We show the scalability and robustness for the hierarchical transformer model.

This part will be covered in Chapter 8 and Chapter 9.

1.8.7 Contribution 7: Construction of 3D abdominal atlases and tissue correspondence modeling

As in human anatomy requires accurate maps, coordinate system and reference space to support effective communication within the region of interest. Body atlas can provide a useful tool to access the relative variability of features. An anatomical template can also help to deal with probabilistic distributions and confidence limits of structure identification, functional variables.

Under the project of the human biomolecular atlas program (HuBMAP), which aims to develop an open and global platform to map healthy cells in the human body. We seek to build image level atlases of the human body. Starting from the kidneys, we study to differentiate and contextualize findings within tissues, organs and systems. We also construct kidney and its substructures, as well as the pancreas atlas. The works define stable spatial reference template to generalize anatomical characteristics across populations. Our main contributions are summarized as:

We constructed the first multi-contrast CT healthy kidney atlas framework for the public usage domain. We proposed a standardized framework optimizing for kidney organs and generalized the anatomical context of kidneys with significant variation of morphological and contrastive characteristics across demographics and imaging protocols. We evaluate the generalizability of the atlas template by transferring the atlas target label to the 13 organs well-annotated CT space with inverse transformation. Unlabeled multi-contrast phase CT cohort is used to compute average and variance mapping to demonstrate the effectiveness and stability of the proposed atlas framework. Our proposed atlas framework demonstrates a stable transfer ability in both left and right kidneys with

median Dice above 0.8.

In the study of kidney substructures, we undertake concurrent studies to validate the following targets: 1) design the clinical study cohort, 2) develop an automatic renal segmentation method for arterial phase CT scans; 3) examine the repeatability and reproducibility of automatic and manual segmentation by three independent observers; 4) estimate the feasibility of using the automatic method for volumetric analysis.

This part will be covered in Chapter 10.

1.8.8 Contribution 8: Application of Quantitative Imaging Biomarkers

Investigations in the imaging structures has been used in clinical biomarker associated with disease, infection and cancer. We have explored quantitative estimations of biomarkers of clinical interest in molecular, histology, radiography characterization of abdominal anatomies such as splenomegaly. Unique challenges emerge when validating machine learning generated imaging biomarkers. We provide spleen volumetric changes for evaluating significant metrics among patients.

This part will be covered in Chapter 11.

CHAPTER 2

Body Part Regression with Self-Supervision

2.1 Introduction

Clinically acquired CT scans can exhibit large variations in field of view, especially for coverage of the chest, abdomen or pelvis (Figure . 2.1). Without pre-processing, such scans are difficult to use for medical image analysis due to lack of spatial consistency. Using pre-processing to remove inconsistency in field of view helps to localize anatomical regions in each body part and enable more precise image registration and machine learning. Recently, Zhang et al. [113] suggested that image quality, quantitative analysis, and anatomical structure localization could be combined to estimate spatial consistency across the human body in CT. The potential applications include content navigation [32], lesion detection [33], classification [34] and segmentation [24, 35], which universally benefit from accurate quantitative assessment of body parts in regions of 1) shoulder and lung, 2) abdomen and 3) pelvis shown in Figure. 2.1.

To achieve slice-wise tissue navigation and to quantify body consistency, the body part regression technique was proposed, which estimated a uniform spatial location (i.e., global position scores in Figure. 2.2) of axial slices for a particular subject [113]. Initially, body part regression was formed as a supervised learning task using deep learning [113]. However, intensive manual annotation is required to prepare the large-scale training cohort. To alleviate the manual efforts, Yan et al. [31] proposed an unsupervised regression network (URN) to perform body part regression in an annotation-free manner. The method required the creation of distance metrics, varied combinations of scalars, isotropic variance, and slice thickness, which resulted in difficulties in creating consistent performance for capturing continuities between slices in medical image volumes.

Challenges and limitations with annotation-free method restrict the generalizability and robustness of models. Therefore, to leverage the current framework, we propose a self-supervised approach named blind-unsupervised-supervision network (BUSN) using robust regression [114] and uses the corrected predictions to provide extra supervision. Our contributions are in four folds: (1) an self-supervised solution to boost current body part regression are designed using 1030 multi-

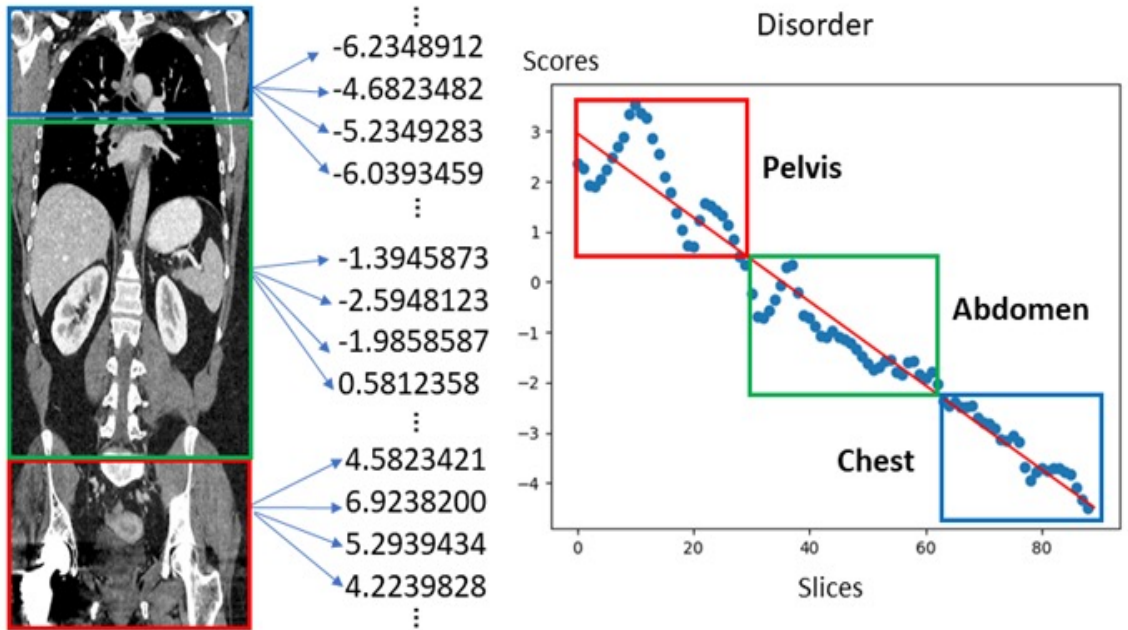


Figure 2.1: Slice disorder problem in three regions with the unsupervised regression network (URN). The left panel indicates the global location scores along slices indices. The body part regression values (blue dots) are inconsistent in the right panel compared with an ideal linear relationship (red line).

center CT scans without using manual labels; (2) a novel unsupervised-supervision method is introduced to achieve robust body part regression; (3) we propose a neighbor message passing correction method to further improve the BUSN results by modeling the spatial relationships between axial slices; and (4) a preprocessing pipeline is proposed using BUSN to normalize the CT volumes spatially, which is evaluated by organ navigation and 3D multi-organ segmentation on normalized scans.

Herein, 1030 whole body CT scans without manual annotation are used to train and evaluate the proposed method. First, regression scores are evaluated with R-squared metric. For body part regression results, the proposed method achieved superior performance compared with baseline methods. Second, organ-wise navigation is performed according to slice by slice scores. This experiment shows that the robust body part regression is able to be used for localization of anatomies. Last, as an example of application, we trained a 3D multi-organ segmentation model for evaluating performance on abdomen CT scans using BUSN as a preprocessing stage.

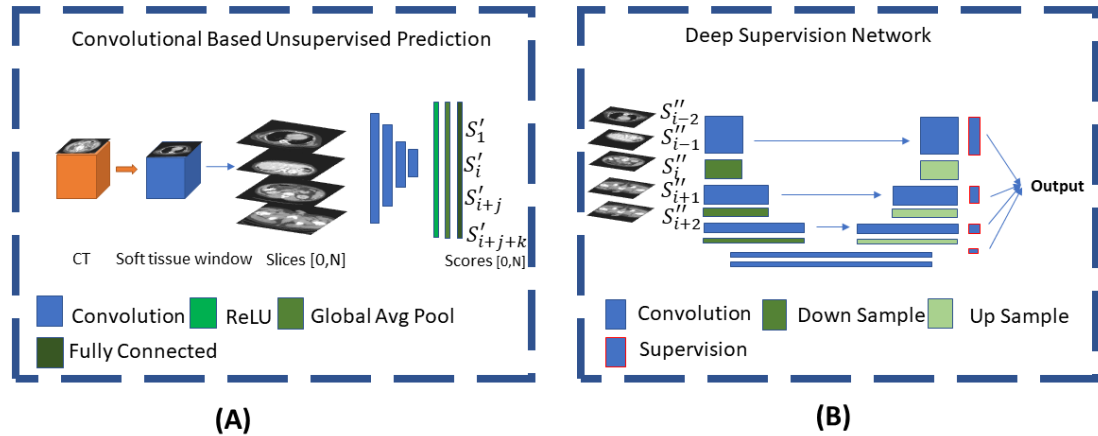


Figure 2.2: Proposed deep blind unsupervised-supervision network (BUSN). Panel (A) is the unsupervised network with robust regression refinement. Panel (B) is the deep supervised network using the refined prediction scores. After training, only the right panel (B) is required to perform body part regression.

2.2 Data

Body part regression: The method is evaluated using a large -scale of 1030 whole body CT scans from multi-center datasets (Tab. 2.1) to compensate the insufficient of views in lung and pelvis. Five multi-center datasets are used only for training, while the sixth, 100 scans of (BTCV) [19] are used for external validation.

Organ navigation: We used the independent 100 whole abdominal CT volumes from BTCV as the external evaluation. We used all 100 research-controlled cases for evaluation. The in-plane resolution ranges from 0.59×0.59 to 0.98×0.98 mm².

Multi-organ segmentation: We used the same external cohorts of BTCV 100 scans, each with all 12 labeled organs, in the multi-organ segmentation task. We integrate the body part regression method as a preprocessing step, where each slice is assigned a slice score. For each scan, the axial slices with score between -6 and 5 are kept in the final 3D volume. The axial slices outside the range will be cropped out. While the zero-padding is applied to fill the volume if the score range is not be able to cover -6 to 5. Last, all scans are resampled to a unified dimension of $[168, 168, 64]$, with the resolution of $2 \times 2 \times 6$ mm for training a 3D segmentation network [115].

2.3 Experiment

2.3.1 Body Part Regression

To evaluate the accuracy of regression scores, we apply BUSN method to predict a score corresponding to each 2D slice. Slices are soft-tissue windowed and fed into the unsupervised network Figure. 2.2. A rough score is predicted according to slice thickness and continuity nature. Then, robust regression refinement is implemented for fixing incorrect scores. Finally, an accurate score is predicted from an encoder-decoder structure with deep supervision. In this section, the deep supervision network is composed of multi-level convolution layers, which are deployed in a symmetric scheme to enable efficient inference. ReLU and max-pooling layers are implemented in the encoder part of the network. Pooling layers helped to enlarge the receptive field of neurons where more contextual information is considered in layers. The decoder part consists of convolution, upsampling and ReLU layers. The convolution filters are set to 3x3x3, while the maxpooling kernel is 2x2x2. The stride is 1 and downsample/upsample factor is set to 2 in each dimension. We used the Adam optimization with learning rate of 0.0001. The batch size is set to 4. All weights are trained from scratch with random initialization. The results are evaluated with R-squared measurement, relative to true anatomical position. The unsupervised BPR method pipeline was employed as state-of-the-art performance for body part regression. To assess ablation performance, the self-supervision method was performed to the target CT scans with robust correction and the BUSN with neighbor message passing scheme (NMP). All experiments are implemented with same data configuration.

The encoder part of the unsupervised learning is learnt from the relative locations and distances between slices. Let L_a denotes the loss given sequential slices.

$$L_a = - \sum_{i=1}^{m-1} \log(S(f(i+1) - f(i))) \quad (2.1)$$

where S is a sigmoid function. $f(i)$ is the predicted score of the i^{th} slice. Let L_b be the correlation between slices.

$$L_b = - \sum_{i=1}^{m-2} |f(i+2) - 2f(i+1) + f(i)| \quad (2.2)$$

The loss function is given by: $L_U = w_a L_a + w_b L_b$. The weights w_a and w_b were empirically set as 1 and 10 according to [31]. L_b indicates the numeric difference between two slice which are proportional to physical distance of two images. Equation 1 keeps the qualitative order of the

regressed slice scores. The value variance between two slice scores is close to the physical distance between the two images. Since we used the sets of neighboring equidistant slices (e.g., slices $j, j+k, j+2k, \dots$), the slice scores should be equidistant as well. In the experiment, the order loss (Eq.1) and distance loss (Eq.2) collaborate to constrain each slice score $f(i)$ towards the direction relative to other neighboring slices. Under the defined loss function (Eq.1, Eq.2), the URN output scores range in -15 and 15. The learned regression scores and patient anatomical body parts correspond well (-15: upper chest, -5: upper liver, 0: lower abdomen, 5: lower pelvis). URN is also robust to the varying position, size and imaging variance.

2.3.2 Robust Regression Refinement

Robust regression [116] refinement is introduced to further correct the inconsistent prediction values (in Figure. 2.2) by hypothesizing that the distribution of the body part regression scores follow a linear distribution. We adopted the random sample consensus (RANSAC) [116] algorithm. As shown in Figure. 2.4, the RANSAC robust regression help removed the outlier predictions (yellow points) in a volume, and forms the corrected pseudo-labels following the evenly distributed scores. The RANSAC is an iterative approach to evaluate parameters from discrete observed data contains inliers and outliers when outliers are presented to be no influence on the values of the evaluation. RANSAC estimates parameters with high degree of accuracy even with large number of outliers. The linear distributed pseudo labels are used for the second stage training. Unlike many robust estimation approaches such as statistics of least-median and M-estimators squares [117] prevailed in image processing, RANSAC was created by resampling technique that presents candidates by minimum number of observations. The aim of the approach is to model the hidden linear trend from heterogeneous input data using robust linear regression (i.e., resilient to outliers). In our context, the robust regression is deployed to correct the discontinuity of the predicted scores as the training labels for the unsupervised learning Figure. 2.2. The robust regression is critical as it helps achieve the expected linearity of score across slice indices.

2.3.3 Deep Supervision Network

The section of deep supervision network uses a symmetric convolutional architecture. As shown in Figure. 2.2, the end-to-end network is defined with multichannel inputs, where each channel is

corresponding to a score number. According to the continuous property of slices, the supervision task is formulated as multichannel pixel-wised image to score regression.

The deep supervision network [118] introduced in the BUSN method Figure. 2.2 is a standard encoder-decoder CNN. The end-to-end network is defined with multichannel inputs (2.5D) to leverage the performance from using a single input slice (2D) every time. The U-Net [65] is used as the backbone. Different from the standard U-Net, we concatenate deep supervision (Figure. 2.2) from different levels to integrate the deep features from coarse to fine. A dense layer is used to convert the long dimensional 1D features from deep supervision to regress a single location score [119, 120]. During the training, we employ the L1 loss for each channel, which is more resilient to outliers compared with L2 loss. To leverage the knowledge of each level and to alleviate the outliers, the deep supervision scheme acquires weighted sum of losses from each level. The final supervision loss L_S is given by

$$L_S = - \sum_{s=1}^N w_s L_s \quad (2.3)$$

where L_s is the loss function of level s , N is the total number of levels, w_s are weight parameters of level s . The weight of final output is set as 1, while the weights of sub-labels are set as 0.1 and 0.01 [118]. N is empirically set to four to present four transpose layers.

2.4 Results

Figure. 2.3 presents qualitative results of body part regression on a randomly selected scan. The URN indicates a slice from lung area to abdomen, a kidney area slice to liver region or an upper pelvis slice to lower pelvis. BUSN and NMP helped fix the ordering problem from the URN model. Figure. 2.4 compares the R-squared error of all methods including URN, BUSN and BUSN with neighbor message passing. Paired t-test is presented as the statistical analysis. From the results, BUSN presents better linearity to unsupervised approach with significant improvement. Figure. 2.3 compares the quantitative results of the body part regression using conventional linear regression and the robust regression of the same cohort. The pure BUSN network fixes the disordering problem in chest, abdomen and pelvis regions using self-supervision.

Across individuals, the body part regression scores should help localize organs. We performed organ-wise comparisons across 100 individuals (Figure. 2.5). The horizontal range indicates the

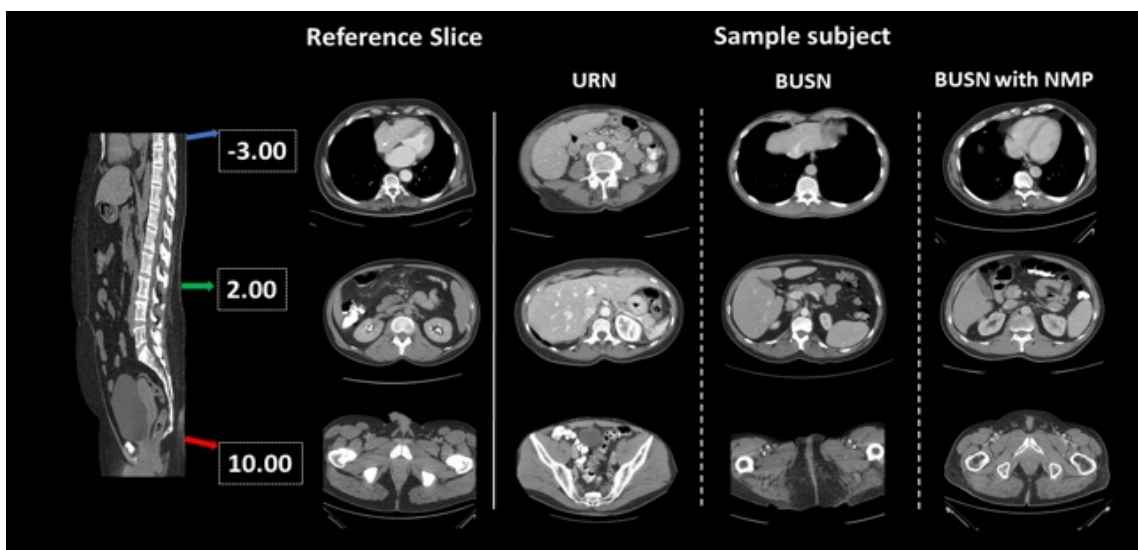


Figure 2.3: The three rows show slices in chest, abdomen and pelvis regions in the same subject, under same regression score (-3, 2 and 10) with four columns (URN, BUSN, BUSN with Neighbor Message Passing and ground truth). The slice predicted by BUSN with NMP is closer to the reference slice.



Figure 2.4: A representative subject was evaluated with URN (left) and BUSN (right). Green scatters are inliers of influence to the regression, yellow scatters are outliers of no influence to the distributed data. Darker blue line indicates the normal linear regression on scatters points, lighter blue line is the RANSAC regressor result according to inliers. Left panel presents the single URN regression with amounts of outliers result in failure of linearity nature in chest and pelvis regions. Right panel shows the testing result of BUSN method, the distributed scores follows good linearity in chest, abdomen and pelvis regions in CT scan. In summary, BUSN takes advantage of self-supervised network, which presents better continuity in regression result among neighbor slices and shows scatter plots without number of outliers.

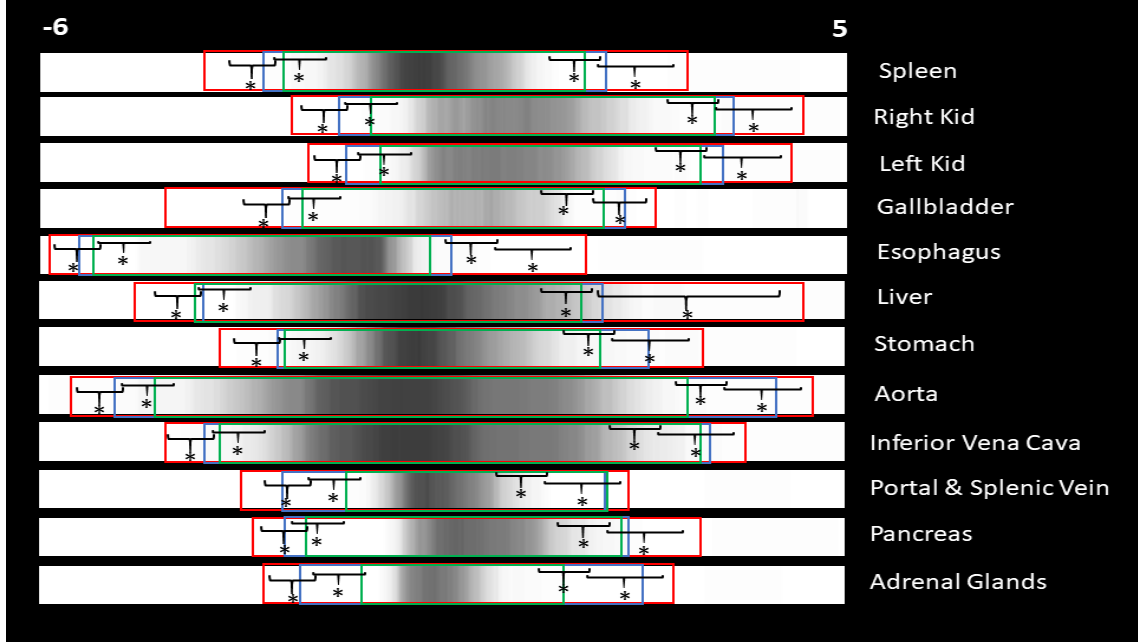


Figure 2.5: Organ navigation task and organ-wise body part regression analysis: Density maps represents the distribution of each organ in whole-body CT scan. The red box range represent the URN method, while the blue box is the BUSN-plain method, and the green box shows the result in BUSN with neighbor analysis. “*” indicates statistically significant (p-value < 0.01 from paired t-test).

distribution of each organ over this cohort. For example, in the URN pipeline, the mean of top border score (Figure. 2.5) of spleen is -4.0217, mean of bottom border is 4.2924, while the means of BUSN is -3.7872 and 3.0158 respectively. A larger range indicates the larger uncertainty in organ localization. The green box shows a more precise evaluation after implemented the proposed BUSN method, which exclude outliers on the boundary of organs. The upper bound of aorta and lower bound of left, right kidney is defined as the abdomen. We empirically selected the scalar -6 to 5 as the range of abdomen region according to the corresponding ground truth label of all of training subjects (e.g. the top most slice that contains esophagus or liver label is score around -5.5, -6 is selected to ensure slices above the top most labeled slice). Table. 2.1 shows the mean Dice similarity coefficient (DSC) and standard deviation of multi-organ segmentation. We compared four methods: (1) without using body part regression, (2) using URN, (3) using the proposed BUSN method, and (4) the proposed BUSN with NMP. Without body part regression, the performance is inferior compared with the results with body part regression method. The average DSC of BUSN with NMP is 0.8145 against URN (0.7991) The BUSN with neighbor message passing performs

Table 2.1: Multi-organ segmentation results with 3D U-Net and different preprocessing strategies are presented with average Dice coefficients. The best performance results marked as bold. BR means bodypart regression. “*” indicates statistically significant (p-value < 0.01 paired t-test) between left and right mean DSC.

Organ	NO BR (Baseline)	URN (Ke et al.)		BUSN (Ours)		BUSN + NMP (Ours)
1.Spleen	92.82 ± 2.13	94.57 ± 2.24	*	94.97 ± 2.05	*	95.61 ± 2.01
2.Right Kid	89.96 ± 2.54	90.41 ± 2.63	*	92.32 ± 2.43	*	93.21 ± 2.17
3.Left Kid	88.93 ± 2.01	90.44 ± 3.21	*	91.26 ± 2.51	*	92.35 ± 2.12
4.Gallbladder	53.94 ± 20.12	54.01 ± 21.14	*	54.73 ± 22.31	*	55.92 ± 18.94
5.Esophagus	74.81 ± 5.47	75.77 ± 7.53	*	76.78 ± 6.36	*	76.98 ± 6.02
6.Liver	94.58 ± 1.94	95.36 ± 2.15	*	95.73 ± 2.86	*	96.01 ± 1.46
7.Stomach	82.98 ± 4.21	84.01 ± 6.01	*	84.72 ± 4.12	*	85.47 ± 3.75
8.Aorta	90.63 ± 3.45	91.16 ± 3.92	*	91.74 ± 3.05	*	91.95 ± 2.33
9.Inferior vena cava	80.65 ± 3.97	81.46 ± 4.62	*	82.86 ± 3.76	*	82.99 ± 2.19
10.Potal&splenic vein	62.17 ± 13.48	63.79 ± 14.27	*	69.01 ± 9.57	*	71.62 ± 9.47
11.Pancreas	67.65 ± 10.27	71.71 ± 10.85	*	73.06 ± 9.48	*	74.21 ± 9.01
12Adrenal glands	63.21 ± 11.36	64.02 ± 13.10	*	64.89 ± 10.84	*	65.17 ± 8.23

achieve the generally highest DSCs of organs with smaller variances. The p-value with paired t-test between No BR and URN is 0.00097, 0.015 between URN and BUSN, and 0.0017 between BUSN and BUSN with neighbor message passing.

2.5 Discussion

3D medical images are intrinsically spatialized, and the location of anatomies/organs are relatively structured. The goal in this study is to predict a continuous, uniformly distributed score for each axial CT slice along with body coordinate values. As the CT slice index in a 3D volume increases, the predicted coordinate scores should be larger dependent. However, the variation of imaging process, position, size lead to morphology difference in human bodies. Herein, we take the full image of each slice, and intends to preserve the spatial context. The pixel-wised feature maps provide strong prior information, then we use the global average pooling and linear layer to obtain the final regression score. Overall, the numeric difference of the predicted slice scores can be approximately corresponded to the spatial context.

In this paper, we propose an unsupervised-supervision learning body part regression framework in a self-supervised paradigm that achieves superior performance without using manual annotations.

Our method outperforms the state-of-the-art methods in terms of the body part regression accuracy. The integration of robust regression analysis leads to the pseudo-ground truth data that are exploited in the context of supervised networks. Furthermore, the part regression enables the superior content navigation and volumetric segmentation. In the segmentation task, we compared our method with the current state-of-the-art performance in the challenge dataset [44]. The averaged DSC score in [44] is 0.832 compared to 0.8179 (ours BUSN). We achieved lower but comparable DSC scores with single 3D UNet model without extra data.

One limitation of the deep learning-based segmentation is the generalizability across different reconstructed scans. In this work, all experiments and results are performed on axial aligned images. So, the model cannot be applied to coronal view and sagittal view slides. Next, -6 to 5 are empirically set to as the range when normalizing different scans as preprocessing according to the experiments. However, body part regression might not be the optimal choice when applying the modal to the CT including the extremities. A detailed comparison of body part regression to detection, such as multi-atlas labeling is expected. In the future, to further improve segmentation result, it is worthy to investigate cropping volumes for each organ specifically and fitting cropped regions into segmentation models. In the segmentation task, the body part regression preprocessing takes 30 seconds on average per case, which indicates reasonable time efficiency on implementation.

In this study, the self-supervised BURN method outperforms the baseline methods. In the future, the usage of pseudo-ground truth data can be employed in the context of supervised learning methods. Herein, we could take advantage of the stability in supervised learning as well as the unsupervised nature of entire framework. Additionally, robust statistics would benefit for many approaches as quality assurance (QA) is a promising avenue to regularize information and knowledge. Therefore, the proposed self-boosted networks might be an opportunity for broader tasks with consistent performance in regression, segmentation, or classification. Investigation into the boosted approach could provide valuable improvements without extra manual efforts.

CHAPTER 3

Anatomy-Aware Semi-Supervised Body Part Regression for Multi-Contrast CT Images

3.1 Introduction

The large-scale medical data enables artificial intelligence (AI) system the capability of modeling general and broad representations across heterogeneous context [68]. Most machine learning algorithms are trained with a well-designed, small data set with a specific region of interest. For instance, computed tomography (CT) for quantifying measurements of body compositions [72] is routinely suggested as the primary screening tool. However, to reduce the radiation exposure, characterization of body depends on the requirements and scanning protocol in each medical institutes [121], resulting acquired CT scans of different field of view (FOV). In addition, there are potential challenges on evaluating variance of single CT slice obtained from longitudinal data (e.g., the Baltimore longitudinal study of aging (BLSA) abdomen and thigh data). This can lead to difficulties of collecting medical image datasets and create accurate AI tool for the clinical contexts.

Quantifying the cross-sectional (slices) images of human body according to spinal anatomy provides a useful detection mean of locating body regions [122]. However, these investigation requires massive manual efforts by clinicians and radiologists and are potential limited by insufficient reproducibility and repeatability (e.g., selecting the consistent slice across longitudinal CTs for measuring body compositions such as fat and muscle).

Another great challenge of training robust body part regression model is the heterogeneity of imaging protocols [121, 123] for enhancement phases. The application of machine learning algorithms relies on the distribution of training samples [68]. There is potential bias of CT enhancement phases for learning body part regression models as portal venous phases data are used solely for constructing the regression task.

Current pre-processing and data curation tools, such as the unsupervised body part regression [124] and BUSN [125], rely on purely self-supervised regularization of physical distances across CT subjects, are insufficient to curate data in an accurate scope. In real clinical practice, clinicians and research experts focus on the anatomy-wise biomarkers for locating meaningful regions

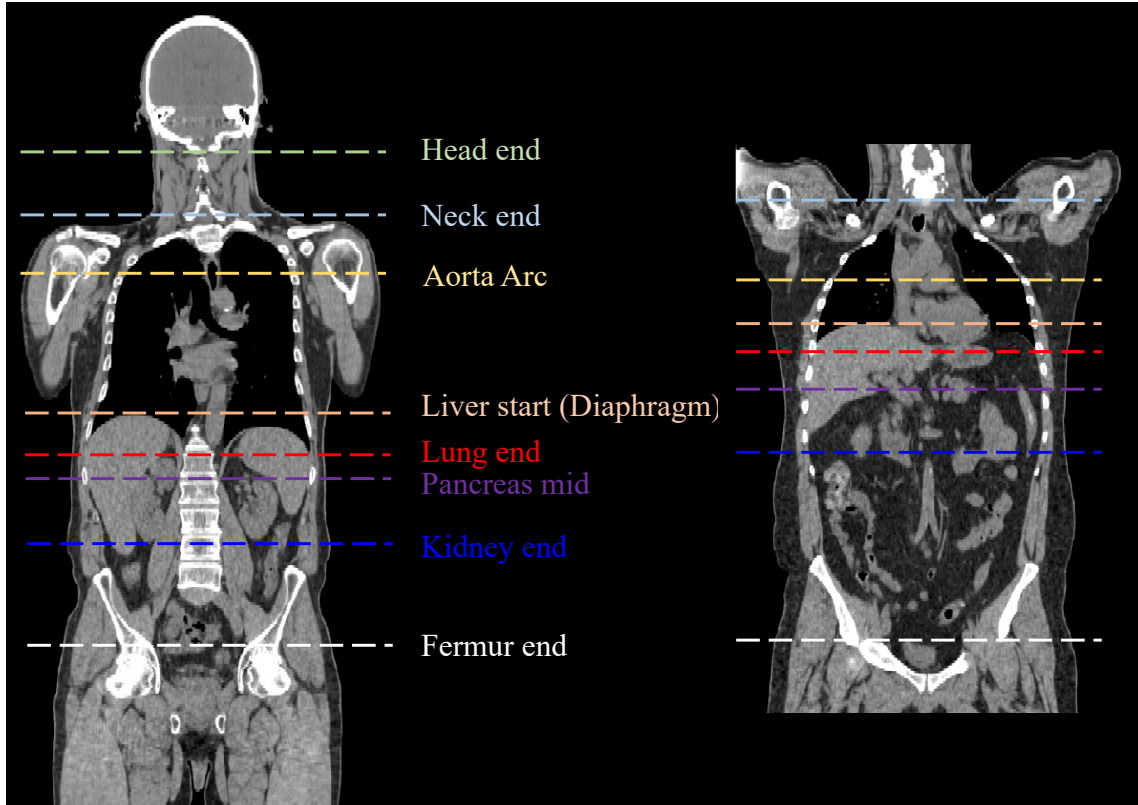


Figure 3.1: Representative physical variance in annotated anatomies' landmarks. The right subject demonstrates shorter distance between target landmarks.

or slices of CT data. For example, the metric of optimal slice for visualizing pancreas tissue can indicate the diabetes patient's phenotype. Recently, transformer-based models demonstrate better representation learning capability when large-scale data presents [88, 126, 127]. The transformer blocks utilize the advantage of self-attention mechanisms show exceptional performance for visual tasks such as detection [89] and classification [88]. Inspired by the good scalability and efficiency of training big transformer models [86], we propose to use vision transformer as the image perception model backbone for our regression task that leveraging 1) large scale medical multi-contrast context; 2) anatomy-aware long-range dependencies for body part regression problem. In this work, we propose to address the current limitations of body part regression systems by an anatomy-aware semi-supervised approach. Regarding above challenges, we pose our contributions as: (1) we introduce a practical approach for reducing body part regression variance; (2) the framework incorporates anatomy landmark context for body part regression. (3) Our method is designed for multiple CT enhancement phases with a large scale of training samples (> 7000 CT volumes), and provides

Table 3.1: Definition of anatomy-based landmarks for body parts. We demonstrate multiple landmarks in abdomen regions for diverse organs such as liver, pancreas, kidneys.

No.	Landmark name	Description	Notation Class
1	Femur end	Slice after the femur bone ends	1
2	Kidney end	Last slice where any kidney can be seen	2
3	Pancreas mid	Slice where pancreas can be largely seen well	3
4	Lung end	Slice after end of lung	4
5	Liver start (Diaphragm)	Slice where liver tissue can be seen	5
6	Aorta Arch	Slice where oval artery arc can be seen	6
7	Neck end	Last slice where neck can be seen	7
8	Head end	Last slice where bone and brain can be seen	8
	Above		9

evaluation based on each contrast phase test cohort; 4) we manually labeled 600 CT volumes for regularizing its spatial context consistency.

3.2 Methods

3.2.1 Localization by Anatomy Landmarks

To investigate the localization performance by anatomies, we manually annotate several landmarks across human body. To avoid bias and region shift of intra-subjects, we label CT landmark slices by its target tissue position instead of vertebrae levels. Table 3.1 introduces description of each landmark, we manually investigate 9 key locations including the end of head, the end of neck, the aorta arch, the start of liver, the end of lung, the optimal slice of pancreas, the end of kidney, and the femur end. Particularly, we annotate several abdomen landmark to increase the requirement of dense and accurate estimation for multiple overlapped organs. The order of the landmarks are not restricted as organs or tissues can be shifted in different subject. For instance, the optimal pancreas slice can appear ahead and behind kidney slices. Figure 3.1 demonstrates two examples of annotated landmarks. It shows the variance between subjects, and it's challenging to differentiate accurate localization only based on physical distances.

3.2.2 Multi-Contrast Dataset Selection

The body part regression model requires data of diverse ROI, we collect a large-scale dataset by several cohorts considering contextual information in the CT volumes. We select different body

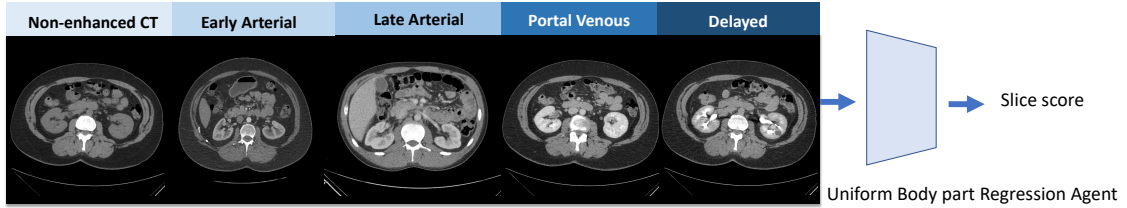


Figure 3.2: The demonstration of multiple contrast enhancement phases data in CT. The arterial phases highlight the aorta, portal venous phase data reaches an optimal brightness of abdomen organs such as liver and spleen, the delayed phase is better at capturing ureter system.

regions such as head, neck, lung, abdomen, and pelvis to model body compositions for a general-purpose representation learning. The selected datasets contains multi-institutes clinical acquired data for capturing heterogeneous information and domain gap. Lastly, to avoid bias to a specific CT enhancement phases, we incorporate five contrast phase CT volumes collected from ImageVU. Specifically, the non-contrast, early arterial, late arterial, portal venous and delayed phase CT scans are obtained from queried clinical cohorts as shown in Figure 3.2. The data cohorts for this work consist 1) ImageVU CT: 4893 subjects, 1,084,114 slices; 2) TCIA Covid 19: 650 subjects, 46,729 slices; 3) HNSCC: 687 subjects, 184,370 slices; 4) NLST: 874 subjects, 137,581 slices. In total, there are 7104 CT volumes covering abdomen-pelvis, thoracis, head&neck regions. The data are clinically-acquired under IRB approval.

3.2.3 Data Pre-Processing

Following [124, 125], we resize all slices to 128×128 , the pixel spacing and slice thickness are maintained originally as the information is crucial for the distance loss. The intensity Hounsfield Unit (HU) is clip to -1000 to 1000 and scaled to 0 to 1. Data augmentation of flip and rotation of probability of 20% are applied to increase the learning capability of anatomy patterns.

3.2.4 Transformer-based MixMatch Model

In this section, we demonstrate our training approach for the semi-supervise body part regression approach.

Transformer Blocks

We show the overview of the proposed approach in Fig. 3.3. The backbone model utilizes a stack of transformers as the encoder for the regression task. The network consists the 2D transformers

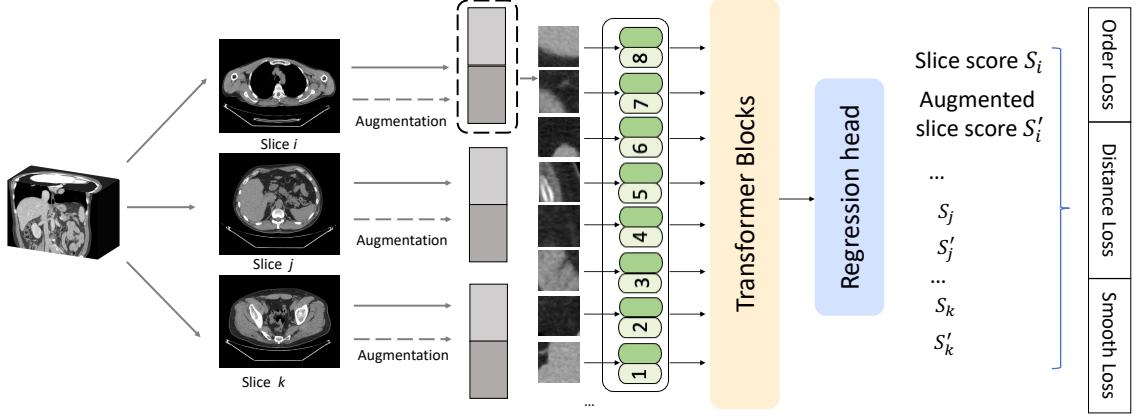


Figure 3.3: The overview of the framework. The volume is sample to several slices, followed by data augmentation for semi-supervised loss. The images are then encoded by patch and position embedding layers.

that exploiting self-attention mechanism. As commonly used in NLP, the transformer blocks take on 1D sequence of input embeddings. In our model, the sequence of a 2D image is embedded $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ with dimension of (H, W) and C is the input channels. Then divide sequences into flattened uniform non-overlapping patches $\mathbf{x}_v \in \mathbb{R}^{N \times (P^2 \cdot C)}$ where (P, P) are the dimension of each sequence and $N = (H \times W)/P^2$ is the length. The linear projection of the image input can be formulates as follow:

$$\mathbf{z}_0 = [\mathbf{x}_v^1 \mathbf{E}; \mathbf{x}_v^2 \mathbf{E}; \dots; \mathbf{x}_v^N \mathbf{E}] + \mathbf{E}_{pos}, \quad (3.1)$$

The position embedding E_{pos} uses the learnable parameter for memorizing the relative 2D grid location of each embedded sequence: $\mathbf{E}_{pos} \in \mathbb{R}^{N \times K}$ to the projected embedding $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times K}$.

After the embedding layer, we utilize a stack of transformer blocks [88, 128] comprising of the multi-headed self-attention (MHSA) and multilayer perceptron (MLP) by:

$$\mathbf{z}'_i = \text{MHSA}(\text{Norm}(\mathbf{z}_{i-1})) + \mathbf{z}_{i-1}, \quad i = 1 \dots L, \quad (3.2)$$

$$\mathbf{z}_i = \text{MLP}(\text{Norm}(\mathbf{z}'_i)) + \mathbf{z}'_i, \quad i = 1 \dots L, \quad (3.3)$$

where Norm is the layer normalization. i denotes the block identifier, while L is the number of transformer layers.

MixMatch Training

The entire framework is shown in Figure. 3.3. In order to incorporate the labeled and large-scale unlabeled data, we employ the holistic semi-supervised learning strategy for training our model. The MixMatch [129] is a dominant self-supervised learning approach that takes advantages of data augmentation and label guessing for additional regularization methods. Given the batch of X as the labeled images with a set of landmark annotations, and the U as the unlabeled CT volumes, the MixMatch provides a set of pseudo labels that regularized by the augmentation pairs. Formally, we employ the $L1$ loss as the regression energy function, the MixMatch is defined as:

$$L_U = \frac{1}{N_U} \sum_{u_j \in U} |y_u - y_{u'}|, \quad (3.4)$$

where N_U is the total number of samples in a batch, u_j is the j th sample of U , u' is the augmentations of sample u . The total semi-supervised training loss is:

$$L = L_X + \lambda L_U, \quad (3.5)$$

Loss Functions

Following the definition in the self-supervised body part regression [124, 125], we use the two objective function for modeling the ordering and distance as the regularization. The order loss is designed for maintain the slice score increasing monotonically as follow:

$$L_{order} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{m-1} \sum_{j=1}^{m-1} L_{order}(\Delta_{s_{ij}}), \Delta_{s_{ij}} = s_{ij+1} - s_{ij}, \quad (3.6)$$

where s_{ij} is the score of j^{th} slice in the volume i .

The distance loss models the numeric difference between two slice scores that are proportional to the actual physical distance in the body. The distance loss is define as:

$$L_{distance} = \sum_{i=1}^N \sum_{j=1}^{m-2} f(\Delta_{i,j+2} - \Delta_{i,j+1}), \Delta_{i,j} = s_{i,j} - s_{i,j-1}, \quad (3.7)$$

where f is the $L1$ loss. The total loss function is combined with the MixMatch loss, order loss and distance loss.

Table 3.2: The quantitative results comparison with baseline methods. The absolute error and scale percentage error are shown according to the 8 manual defined landmarks.

No.	Landmark	UBR		BUSN		SemiBR		SemiBR+Smoothing	
		Ab Err	Scale Perc	Ab Err	Scale Perc	Ab Err	Scale Perc	Ab Err	Scale Perc
1	Femur end	1.672	5.57%	1.149	3.83%	0.275	0.91%	0.253	0.84%
2	Kidney end	2.124	7.08%	1.694	5.65%	0.601	2.00%	0.582	1.94%
3	Pancreas mid	2.261	7.54%	2.071	6.90%	0.674	2.24%	0.601	2.00%
4	Lung end	1.834	6.11%	1.637	5.46%	0.527	1.76%	0.497	1.66%
5	Liver start	1.682	5.61%	1.407	4.69%	0.205	0.68%	0.198	0.66%
6	Aorta Arch	2.474	8.25%	2.027	6.76%	0.633	2.11%	0.574	1.91%
7	Neck end	1.407	4.69%	1.196	3.98%	0.204	0.68%	0.203	0.68%
8	Head end	1.637	5.46%	1.334	4.45%	0.316	1.05%	0.287	0.96%
Avg	Landmark set	1.886	6.29%	1.564	5.21%	0.429	1.43%	0.399	1.32%

3.3 Experiments

We conduct the experiment and evaluate the effectiveness of our proposed approach on the manual labeled data. Our study compares UBR [124] and BUSN [125] as baselines in terms of statistical measurements. We also add the robust smoothing strategy as an additional post-processing step to compare the our methods.

Implementation Details

For defining body part regression score distribution, we normalize the score into -15 to 15, the larger number indicates upper body regions until top of the head. We randomly split 20% the labeled CT volumes as the held-out test set. The remaining labeled data are then randomly split to five folds for training and validation. The entire unlabeled data are used as training. We apply our pipeline in Pytorch and MONAI using single 16G NVIDIA GTX 6000 GPU. We train the model for 50 epochs. Adam optimizer is used with an initial learning rate at $1 \times e^{-3}$ and decay rate of 0.999. We use 12 block of transformer layers as the model backbone.

Evaluation Metric

We use the entire annotated volume slices as the reference table, which establish a mapping between anatomies and scores in range of -15 to 15. The ground truth label scores are determined by linear distribution which defined by the manual labeled landmarks. We use absolute error and scale percentage difference as the metric for evaluating body part regression scores in volume. The absolute error is simply $|s_{pred} - s_{truth}|$, the scale percentage error is calculated by $\frac{1}{30}|s_{pred} - s_{truth}|$.

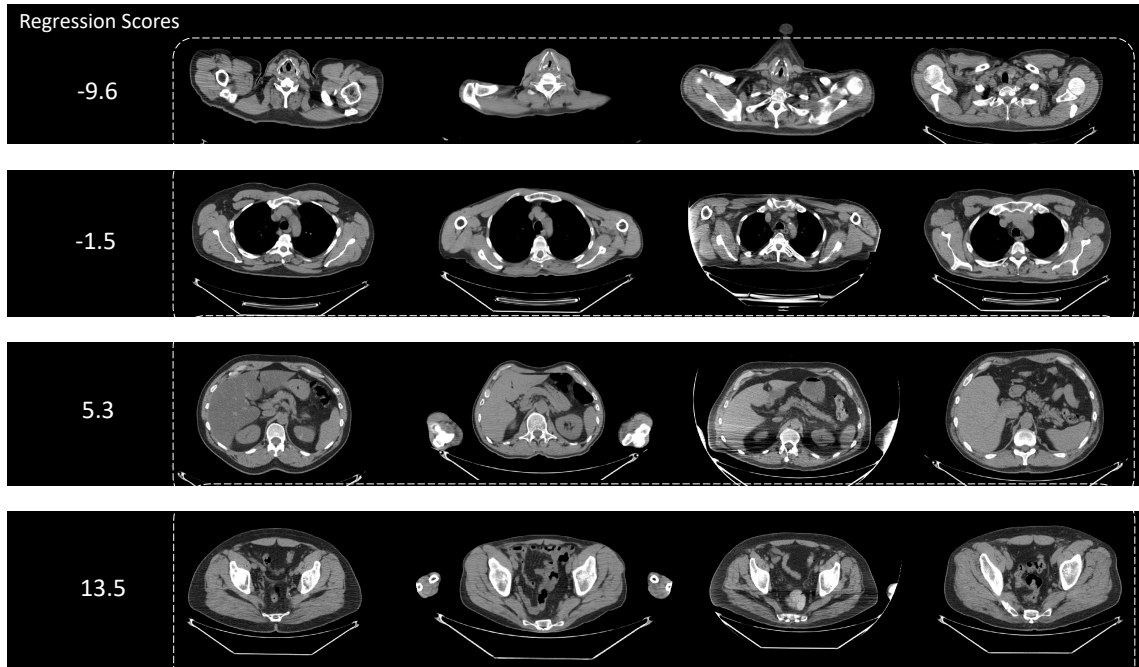


Figure 3.4: Qualitative visualization show the effectiveness of maintaining cross subject consistency of slice scores. A neck, aorta arch, pancreas, and femur end slice for inter-subjects at same predicted scores are shown.

Given absolute error of 1.5 and slice thickness of 2mm as an example, we derive the model can differentiate physical distance error by 3mm.

3.4 Results and Discussion

3.4.1 Anatomies landmark boosted regression accuracy

We conduct experiments comparison between our approach to the baselines in Table. 3.2. The results demonstrate the prediction accuracy with the 8 annotated landmark slice scores including femur end, kidney end, pancreas mid, lung end, liver start, aorta arch, neck end and head end. We observe that our approach surpass baselines and achieves average absolute error with 0.429 against 1.564, 1.886, respectively. The scale percentage rate is 1.43% compared to 5.21% and 6.29% of baselines. The 0.429 indicates the general physical distance error can be controlled within 1mm, given slice thickness around 2mm. In addition, the abdomen slices such as kidney (0.601) and pancreas (0.674) are harder to identify compared to other positions (e.g., femur end: 0.275, neck end: 0.204). This indicates the composition complexity between lung and abdomen regions, compositions such as fat, muscles, organs can be shifted among subjects. Furthermore, we per-

Table 3.3: The testing results categorized by different CT contrast phases. The performance show late arterial and portal venous phase data are of higher accuracy partially due to the larger size of training data.

No.	Landmark	Non-Contrast		Early Arterial		Late Arterial		Portal Venous		Delayed	
		Ab Err	Scale Perc	Ab Err	Scale Perc	Ab Err	Scale Perc	Ab Err	Scale Perc	Ab Err	Scale Perc
1	Femur end	0.319	1.06%	0.263	0.88%	0.259	0.86%	0.240	0.80%	0.334	1.11%
2	Kidney end	0.681	2.27%	0.575	1.92%	0.568	1.89%	0.549	1.83%	0.699	2.33%
3	Pancreas mid	0.632	2.11%	0.593	1.98%	0.531	1.77%	0.507	1.69%	0.681	2.27%
4	Lung end	0.525	1.75%	0.504	1.68%	0.484	1.61%	0.413	1.38%	0.527	0.18%
5	Liver start	0.251	0.84%	0.226	0.75%	0.242	0.81%	0.209	0.70%	0.275	0.92%
6	Aorta Arch	0.619	2.06%	0.581	1.94%	0.501	1.67%	0.497	1.66%	0.620	2.07%
7	Neck end	0.237	0.79%	0.211	0.70%	0.198	0.66%	0.185	0.62%	0.248	0.83%
8	Head end	0.348	1.16%	0.274	0.91%	0.256	0.85%	0.250	0.83%	0.362	1.21%
Avg	Landmark set	0.452	1.51%	0.403	1.34%	0.380	1.27%	0.356	1.18%	0.468	1.56%

Table 3.4: The results show the reproducibility and repeatability of the proposed automatic method and two independent readers. It demonstrate the error between SemiBR and readers are similar to inter-reader assessments.

No.	Landmark	SemiBR-Rater 1		SemiBR – Rater 2		Inter-Rater	
		Ab Err	Scale Perc	Ab Err	Scale Perc	Ab Err	Scale Perc
1	Femur end	0.245	0.85%	0.251	0.84%	0.213	0.71%
2	Kidney end	0.591	1.97%	0.545	1.82%	0.450	1.50%
3	Pancreas mid	0.612	2.04%	0.631	2.10%	0.681	2.27%
4	Lung end	0.501	1.67%	0.456	1.66%	0.485	1.62%
5	Liver start	0.204	0.68%	0.267	1.52%	0.154	0.51%
6	Aorta Arch	0.610	2.03%	0.755	2.52%	0.667	2.23%
7	Neck end	0.298	0.99%	0.263	0.88%	0.385	1.28%
8	Head end	0.349	1.16%	0.392	1.31%	0.403	1.34%
Avg	Landmark set	0.426	1.42%	0.445	1.48%	0.431	1.45%

form the same smoothing post processing using the RANSAC regression algorithm. The robust regression can provide further smooth scores by excluding outlier predictions in a CT volume. The SemiBR+Smoothing achieves the best quantitative result of 0.399 on average.

In Figure. 3.4, we demonstrate the qualitative results on the SemiBR predictions. Four body regions are presented by selecting same regression slice scores. -9.6 indicates the slice near neck end, we show four randomly picked subjects with the score. -1.5 , 5.3 , and 13.5 denote the aorta arch, pancreas and the furmer end regions. The visualization show the regression consistency across different subjects.

3.4.2 Robustness to Multi-Contrast CT Scans

We also show the testing performance according to CT contrast enhancement phases in Table. 3.3. We observe the portal venous phase data are of best performance. This is due to the potential larger cohort of the phase. The late arterial, early arterial phase have the second and third best results with 0.380 and 0.403, respectively. The non-contrast and delayed phase data are of similar contrast except the concentrated dye medium in the ureter system. The performance is 0.452 and 0.468 for non-contrast and delayed phase testing data.

3.4.3 Reproducibility and Repeatability

Table. 3.4 shows the comparison between the proposed deep learning based model and different interpreters. First, the SemiBR prediction is evaluated against the first manual label, the average absolute error is 0.426 and scale percentage error is 1.42%. We also compare the same prediction with the second independent rater's label, the absolute error and scale percentage are 0.445 and 1.48%. To investigate the reproducibility and study whether the automatic method can be repeated for the testing cohorts with manual annotations, we calculate the performance between two raters. The inter-rater absolute error is 0.431 and the scale percentage is 1.45%. We observe that the inter-rater one is similar to the SemiBR-Rater1 assessment, and the SemiBR-Rater2 assessment. This indicates the stability and consistency of the automatic method, and show sufficient reproducibility and repeatability.

3.5 Conclusion

We introduce the SemiBR, a semi-supervised learning approach incorporating the self-supervised tasks. The large-scale collected data is sufficient to provide rich spatial, diverse context of CT volumes. In addition, we address the challenge of the applying body part regression and show its robustness to multi-contrast data. The major results show our approach achieves the best performance that can distinguish physical distance within 1cm. We are interested to exploring this quantitative tool for downstream medical tasks.

CHAPTER 4

Phase Identification for Dynamic CT enhancements with Generative Adversarial Network

4.1 Introduction

Dynamic contrast enhancement Computed Tomography (Dynamic CT) is widely used in the clinical diagnosis [37]. For quantitative measurement in tissues, the standard Hounsfield Units (HU) scale is used in contrast-enhanced CT scans for describing radiodensity. In previous studies, contrast enhanced CT has been shown as an important means for identification of organ physiology, lesions and abnormalities [36, 130, 131]. To capture complex relationships of tissues, enhancements are acquired by injecting iodine contrast materials into peripheral veins [39]. As time proceeds, the medium helps illuminate through aorta, heart atrium, pulmonary vein, lung, liver, spleen, pancreas, kidneys and urinary system respectively [132]. However, the contrast agent timing information is often tagged manually by physicians. Therefore, missing information or mislabeling can happen, which is especially problematic in large-scale studies. Revisiting phase knowledge and contrast protocols in meta-data is resource intensive. Moreover, correcting phase labels is hard and challenging due to variations in tissue, contrast material, injection protocols, vascular dynamics and metabolism. Herein, we explore automatic means for classifying contrast enhancements for application in diagnose pathology.

Deep learning approaches have been widely used for classification tasks [24, 25, 28]. Instead of generating shallow feature components [133], deep models such as VGG [134] and ResNet [24] can encode very deep features by hundreds of convolutional layers. Traditionally, the image classification model is supervised by statistical learning-based loss functions (e.g., margin loss and cross entropy loss [135]). To leverage the classification performance, the intrinsic embedding of the input images are encoded using variational autoencoders [136]. From another perspective, generative adversarial net (GAN) [137, 137, 138] is introduced to the classification tasks using a data-driven discriminator, typically with a generator [139, 140, 141, 142]. More recently, the multi-domain discriminator [143, 144] is presented to solve classification problem and improve image synthesis performance.

In abdomen CT imaging, contrast enhanced scans have been explored for detecting abnormalities [33, 145], identifying tumors [119], and segmenting organs [4]. However, few studies have been explored for contrast phase classification, whose aim is to classify an unlabeled contrast CT slice with correct phase label [40, 41]. In this study, we target classification of five typical contrast enhancement phases: (1) non-contrast (NC), (2) early arterial phase (EAP), (3) late arterial phase (LAP), (4) portal venous phase (PVP), and (5) delayed phase [146]. The examples of different phases are shown Figure. 4.1.

In this paper, we propose a multi-domain contrast disentangling GAN (CD-GAN) to perform dynamic CT phase classification, by learning disentangled representations across contrast phases features. The method (Figure. 4.2) takes 2-D slices to learn an intermediate representation, then reconstruct a synthetic contrast enhanced image. The goal of using generated image is to employ adversarial loss and synthetic classification as data augmentation for improving discriminator’s performance. Our contribution in this study are as follows:

We propose CD-GAN, an adversarial learning network to perform contrast phase classification by learning from both real and synthetic images.

1. A two-step generator is introduced, which learns the disentangled representations and domain specific features among multiple contrast CT.
2. We perform quantitative and qualitative evaluations on of the prevalent baselines (VGG, ResNet-50, StarGAN and 3DSE).

4.2 Materials and Methods

4.2.1 Dataset of Studies

Data retrieval. We performed de-identified CT studies retrieval from the ImageVU project of the Vanderbilt University Medical Center (VUMC) under IRB approval. In the training process, we use 400 subjects (36350 2-D slices). These subjects are independent and unpaired with five contrast phases. Experimental data was separated into 1) 80 non-contrast phase subjects, 2) 83 early arterial phase (EAP) subjects, 3) 67 late arterial phase (LAP) subjects 4) 92 portal venous phase (PVP) subjects, and 5) 78 delayed subjects. The average number of axial slices for all training scans is 91 with variance of 7.2. These 36350 2D slices are obtained from the 3D CT scans. The slice

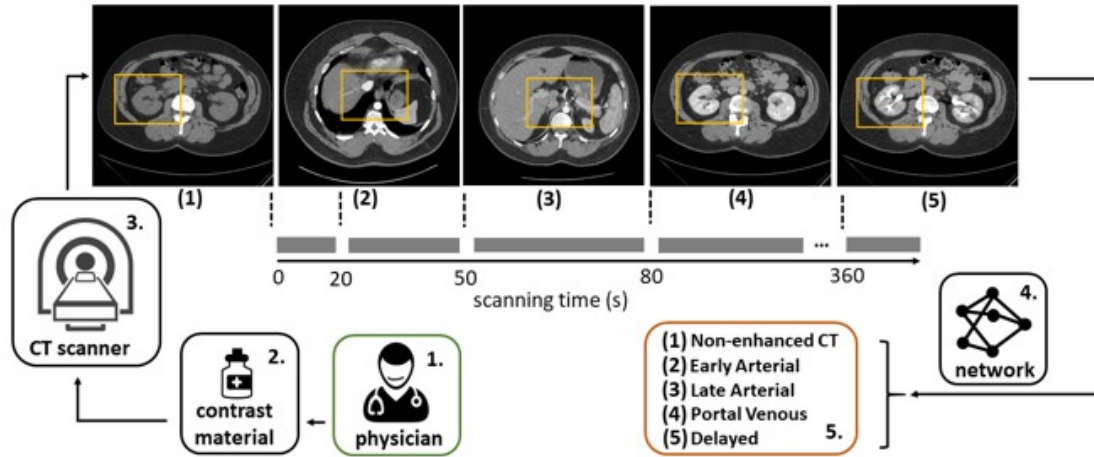


Figure 4.1: Schematic illustration on enhanced CT scan classification. Physician (box 1) assigns contrast materials (box 2) to patients before screening. Different contrast phases scans (1) – (5) are acquired by CT scanner (box 3) depend on time periods (middle axes): 1) non-contrast, 0s or without contrast medium, the box shows the consistent contrast among tissues, 2) early arterial, 15-20s after injection, the optimal illuminated aorta is shown in the box, 3) late arterial, 40-50s after injection, the yellow box shows the light aorta while portal veins starts to be bright, 4) portal venous, 70-80s after injection kidney cortex starts to be bright, and 5) Delayed, 6-10 minutes after injection, the ureter is very bright shown in the box. Last, different contrast unknown scans are fed to our network system (box 4), which outputs a phase label (box 5).

thickness of these CTR volumes varies from 0.1 to 6 mm. In this initial query, we performed 2004 de-identified subjects from medical center. To obtain comprehensive studies on contrast phases, we assessed each subjects' longitudinal study with paired multi-phase scans. Out of 2004, 400 subjects are selected with 2-4 phases' scans. 20 subjects are retrieved with all five phases volumes. The retrieved CT scans are acquired from two manufacturers, Siemens and Philips. Before scan acquisition, contrast agents were injected with amounts varying between 120 and 140 ml. CT signals were acquired at different timing during imaging cycles. The in-plane pixel dimension of these subjects varies from 0.78-0.86 mm. The section thickness ranged from 1.2-3.5 mm. The resulting the number of slices ranged from 80 to 120.

In addition, 20 subjects with paired scans from all five contrast phases (100 scans in total) are retrieved and coded. Scans in a subject are acquired from sessions along with different contrast protocols. The paired 20 studies are used as external testing.

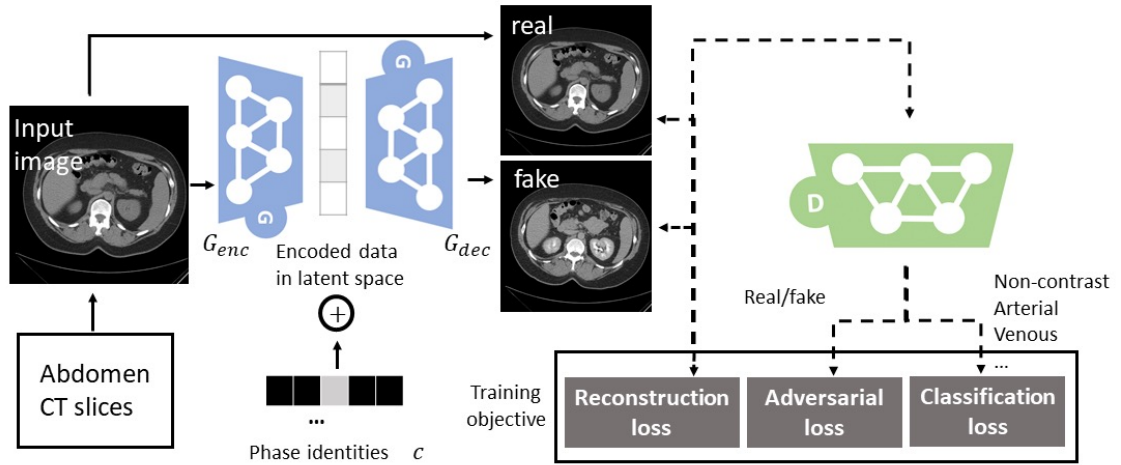


Figure 4.2: The training pipeline of the proposed method. The encoded data in latent space are acquired from the encoder part of generator. The phase identities are concatenated with the intermediate representation before being fed into the decoder part of generator. On the right, the proposed classifier uses both adversarial loss and contrast classification loss on real and synthetic image in the training.

4.2.2 Preprocessing

We convert each body CT scan to axial view 2-D slices with the original resolution of 0.8×0.8 mm and dimension of 512×512 . For consistency, we use a wide HU window from -1000 to 1000 to exclude intensity outliers. Then, image intensities are linearly normalized to -1 to 1. 2D images are resized to 128×128 , we randomly sampled 96×96 region for data augmentation.

Clinically acquired CT scans can exhibit large variance in volume size, so we adopt a critical pre-processing with body part regression [31, 114]. The body part regression helps to remove slices on inconsistency volumes, and to localize anatomical regions automatically. In order to capture contrast feature in abdomen structures, we used the pre-trained model from unsupervised regression network [31] to navigate slices in abdomen region (scalar reference index ranges from -6 to 5).

4.2.3 Contrast Disentangling Generative Adversarial Network (CD-GAN)

The network structure of CD-GAN is shown in Figure. 4.3. GANs are generative models to learn the mapping conditioned on input image x and contrast phase code c . In architecture of a GAN, the generator is trained to predict outputs that fools the discriminator, while the discriminator is trained to distinguish the real and fake images. In our method, the discriminator is not only used to

detect adversarial samples, but also to classify the phase contrast. Our rationale focuses on using adversarial learning to guide the training process, which is to make classification more robust by alleviating spatial inconsistency and leveraging overall performance.

Adversarial Loss: In CD-GAN, the adversarial loss is used to perform adversarial learning [147]. In order to distinguish real and synthetic images and perform the minmax problem, we adopted the conditional GAN objective:

$$L_{adv} = E_{(x,y)}[\log D(x,y)] + E_{(x,c)}[\log(1 - D(G(x,c)))] \quad (4.1)$$

where x is the real input image, G generates the synthetic image $G(x,c)$ which conditioned on input and target contrast phase identity c . D aims to distinguish real, synthetic image and maximize the objective: $\min_G \max_D(L_{adv})$. The discriminator is conditioned on both real image x and synthetic image y .

Contrast Classification Loss in Discriminator. To achieve classification in adversarial learning, we designed an auxiliary classification loss on top of D . The discriminator aims to classify both real and synthetic images to proper phase categories by given target phase identity c' and conditioned phase identity c respectively. For real image path, the contrast classification loss is defined as:

$$L_{cls} = E_{(x,c')}[-\log(1 - D_{cls}(c'|x))] \quad (4.2)$$

where $E_{(x,c')}$ denotes the probability distribution over its true contrast phase labels, and D tries to classify real scan slices to its ground truth identity c' . For synthetic image path, the contrast classification loss is denoted as:

$$L_{cls'} = E_{(x,c)}[-\log(1 - D_{cls}(c|G(x,c)))] \quad (4.3)$$

where c is the assigned phase identity of the synthetic slice. D tries to classify the synthetic image to its assigned identity c . In this work, c is randomly selected for each image.

Contrast Disentangling Generator: By minimizing the adversarial and classification losses, the generator is trained to synthesize images. The synthetic images are simulated with phase knowledge and original contexts. In order to capture variations in CT scans, G_{enc} learns a discriminative

representation by encoding image to feature space: $f(x) = G_{enc}(x)$. To explicitly disentangle the feature representation, the encoder learns variation information to an intermediate representation. Then, G_{dec} synthesizes a CT slice: $x' = G_{dec}(f(x), c)$ with phase identity c . The phase identity c is provided by a hot vector c , with target enhancement phase being 1. The learned intermediate representation $f(x)$ aims to preserve discriminative context while separate contrast identities. With the phase identity added to G_{dec} , the encoder is trained to disentangle the phase variation from $f(x)$. Herein, $f(x)$ is both generative and discriminative. The generator in the model is trained to synthesize images that are close to the given target domain. However, minimizing the adversarial loss and classification loss cannot assure that generated images map the context of input image while editing only the phase-related information. In this study, we employed the cycle consistency loss [137] to the generator:

$$L_{rec} = E_{(x,c)}[||x - G(G(x,c), c')||] \quad (4.4)$$

The reconstruction loss involves the forward-backward cycle. (where the generator takes the original image as input), and also reconstruct the synthesized image back to the original domain. In this experiment, we employed the L1 norm in the cycle reconstruction loss. The discussion of norms can be found in [14].

Network Structure: The architecture of the proposed network is shown in Figure. 4.2. We employ the DR-GAN [138] for G_{enc} . Batch normalization and ReLU are applied after each convolutional layer. The intermediate representation is acquired from the AvgPool layer in the encoder network. Then, the representation is concatenated with phase identity c . In order to perform high resolution image synthesis, the decoder part G_{dec} consists of a series of transpose convolutions that transform the concatenated representation into a synthetic image. The discriminator D is designed on top of PatchGAN [148] with two extra convolutional layers followed by LeakyReLU. The classification result is predicted from fully connected layer with a softmax function in the end of net D .

Full Objective: The GAN architecture achieves a global optimum in the minmax game. In CD-GAN, D and G improves each other during the iterative training. With D achieving better performance of distinguishing real and synthetic image and classifying phase identity, the G is trained for memorizing phase identity features to compete with D . Overall, three objectives are

presented in CD-GAN. First, G_{enc} is trained to disentangle the domain knowledge from contrast enhancement CT slices. By using an assigned phase identity, the G_{dec} is learned to inverse the process that synthesize a contrast preserving image. Last, the discriminator performs phase identity classification on real/ synthetic as adversarial loss converges.

4.2.4 Baseline Architectures

We compared CD-GAN with a series of state-of-the-art approaches, including (1) VGG-16, a convolutional network for classification and detection, (2) ResNet-50, a very deep network for image classification, (3) StarGAN, a multi-domain GAN architecture for capturing feature identities, and (4) 3DSE, a 3D model for phase detection. (1) to (3) are implemented with 2D models, results are evaluated at the scan-level by majority vote. (4) is trained using 3D volumes without a voting algorithm.

VGG Net: VGG-16 is a convolutional neural network proposed for ImageNet [28]. The model consists of a stack of convolutional layers, in which kernels are used with a small receptive field (3 x 3). At the end of the network, VGG-16 includes 1x1 convolution filters, which can be regarded as a linear transformation of input features. Five max-pooling layers are carried out as spatial pooling after convolutional layers, which is performed by 2x2 and stride of 2. Hidden layers are followed by ReLU non-linearity transformations. The Softmax function is used at the end of the network to perform multi-class classification.

ResNet50: The ResNet used residual mapping function to address the degrading problem in very deep networks. ResNet integrated skip connections with convolutional layers as residual blocks. The ResNet-50 consists of 50 convolutional layers with a filter of 3x3, shortcut connections are added to each pair of kernels. For each residual bottleneck block, we use a stack of 3 layers of conv.

StarGAN: StarGAN used a single generator that learns mappings among multiple domains, while discriminator employed classification loss. The generator is a simplified U-Net shape [26], which has two convolutional layers with a stride of 2 and six residual blocks. The decoder part consists of two transposed convolution layers with an up sampling factor of 2. Domain codes are concatenated with images before feeding into the generator, which performs an end-to-end synthesis. The discriminator adopted PatchGAN. Instance normalization is used in the generator and no

normalization is used in the discriminator. The discriminator minimizes only the classification error associated with the known labels.

The 3DSE: 3DSE model consists of two convolutional layers, each followed by a ReLU and pooling layers. The model introduced squeeze and excitation (SE) layers, which increase descriptive capacity and inject global information. 3DSE is implemented as a light-weighted model with a smaller parameter size than typical 3D models [41].

The HU Parameters. To establish the proposed method as an automatic approach to compare with manual procedures, we performed labeling repeatability and reproducibility of manual identifications. We performed a second round of manual labeling; the testing set is interpreted by another observer. The interpreter was instructed to verify the contrast phase slice-by-slice independently. The same procedure was used to label phase stages. The second interpreter (independent observer) perform manual identification without consulting others. The accuracy result is shown in Table. 4.1. We performed manual identification to assess the phase stages. The interpreter was instructed to verify the contrast phase slice-by-slice. The interpreter is supervised by experienced radiologists (> 10 years of experience in radiology). In case of disagreement, consensus meetings were held and performed quality check, adjustment on necessary subjects.

Hounsfield units are representing tissues varied significantly among enhancement phases. We manually labeled each training subject under the supervision of radiologists. HU values for aorta by imaging phases used in terms of magnitude were EAP > LAP > PVP > Delayed CT > non-contrast CT. The HU value for portal vein were PVP > LAP > EAP > Delayed CT > non-contrast CT. We set vessel specific Hounsfield Unit parameters (i.e., aorta in the late arterial phase were found to have central ROI as a fixed HU, i.e., portal venous in the late arterial phase were found to have central ROI as the fixed HU) as an extra validation method, which could be used to quantify visual analysis for identifying studies in the correct phases. In this study, we manually checked HU parameters for all EAP, LAP and PVP scans for all training images, since the non-contrast phase can be detected by identifying the absence of contrast and the delayed phase indicator is the contrast in the ureters. The mean HU value in aorta central ROI for EAP, LAP and PVP of training data are 242.0824.98, 191.4217.26 and 150.5219.73, respectively. The mean HU value in portal vein central ROI for EAP, LAP, PVP of training data are 103.2914.19, 153.1520.42 and 162.8118.18, respectively. We set the lower bound of mean aorta HU in LAP (173.74) as the fixed HU (xHU), the lower bound of mean

portal vein HU in LAP (132.73) as the other fixed HU (yHU). We then use the manual validation as an additional comparison on testing images.

4.2.5 Training Parameters

To provide phase transfer labels, we assign each slice with all five phase identities including the presented phase of input. The phase identity is provided by a hot vector and concatenated with the output of the encoder part of the generator. The generator reconstructs five enhancement phases for each image. We follow the optimization strategy in [20], the batch size is set to be 8. All weights are trained from scratch. The Adam optimizer is used with learning rate of 0.0001 and $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We implement learning rate decay of step 1000 every 100000 iterations. We used step of 1 for alternating between optimizing discriminator and generator [149]. This helps discriminator to maintain optimal solution as the generator converges slowly, which discriminator presents strong supervisions for both real, synthetic images and classification labels.

Training continues with 500,000 iterations and a complete time of about 36 hours on GPU. Implementations are performed using NVIDIA Titan X GPU 12G memory and CUDA 9.0. Training, validation, and testing are executed on a Linux workstation with Intel Xeon CPU, 32GB of RAM. The code of all experiments including baseline methods are implemented in python 3.6 with anaconda3. Networks and frameworks are implemented in Pytorch 1.0.

4.2.6 Experimental Design

We conducted experiments on five phase classifications to evaluate the effectiveness of CD-GAN. First, we compared baseline methods with CD-GAN training on unpaired CT scans. To set the hyper parameters, we implemented standard five-fold cross validation, 400 scans are split into five complementary folds, each of which contains 80 cases. For each fold of evaluation, we use 4 folds as training and validation on the remaining fold. Then, in order to prove stability and sensitivity, we test trained models on the withheld paired subjects. Experimental results are assessed by accuracy score across five phase labels. The claim of statistical significance is evaluated by Stuart-Maxwell test for multi-class confusion matrix ($p - value < 0.01$). In terms of evaluation of volume-wise performance across 2D methods, the scan-level result is obtained from the same fusion strategy of majority voting among all 2D algorithms.

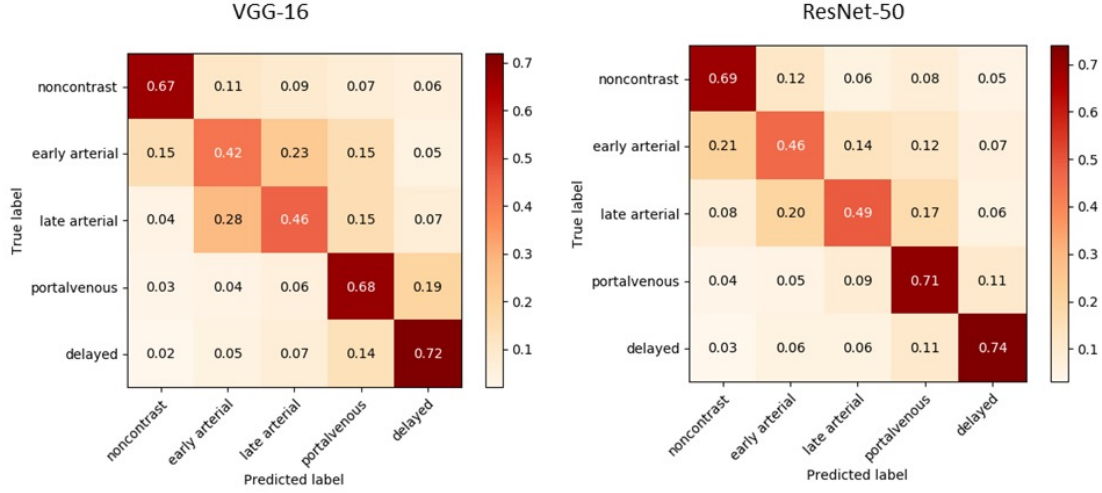


Figure 4.3: Scan-level confusion matrix result on withheld 20 paired subjects of VGG backbone networks. ResNet-50 (right) achieves higher accuracy than VGG-16 (left) by using stacked convolutional layers and skip connection design.

4.2.7 Evaluation Metrics

We evaluated multi-class classification performance with standard normalized confusion matrix. Confusion matrix is a table layout the performance of classification tasks. Each row of the table represents the instances in the predicted class, each column represents the instances in an accrual category. Experimental results are measured by accuracy score.

$$accuracy(p_i, p'_i) = \frac{1}{N} \sum_{i=0}^{N-1} 1(p_i, p'_i) \quad (4.5)$$

where p is the predicted value, p' denotes the corresponding real score, N represents the total number of samples. The fraction of prediction values over N is defined as the classification accuracy.

4.3 RESULTS

4.3.1 Quantitative Evaluation

Figure. 4.3 and 4.4 present the scan-level quantitative results on five enhancement phases. The confusion matrix shows the accuracy of CD-GAN and baseline methods. CD-GAN achieved superior performance over the baseline model on the ImageVU dataset. Compared to both the conventional synonym models (VGG-16 and ResNet-50) and the approach with GAN architectures, our proposed method employed advantages of GAN jointly with classification training achieves superior

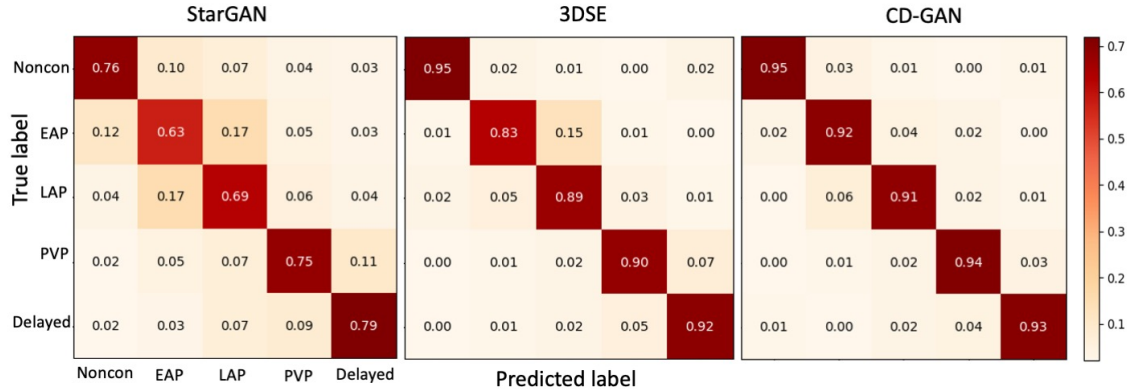


Figure 4.4: Quantitative result on withheld 20 paired subject of GAN architectures and 3DSE. CD-GAN achieves consistent superior performance than StarGAN and 3DSE. Order of labels are the same as Figure. 4.3, true labels top-down, predicted labels left-right are noncontrast, early arterial, late arterial, portal venous and delayed respectively.

results. Figure. 4.3 reports accuracy score on VGG-16 and ResNet-50. ResNet-50 uses VGG net as backbone but deeper with add-on skip connections. The ResNet-50 achieves consistently higher accuracy, which indicates deeper feature caption with residual blocks leverages the performance of phase detection. We speculate that the ResNet-50 stacks more convolutional layers can adapt phase identity more accurately and provide global feature information, which is beneficial in the enhancement detection contexts.

Figure. 4.4 compares methods with GAN architectures and state-of-the-art model 3DSE. Compare to VGG backbone networks in Figure. 4.3, StarGAN achieves significant improvements $p\text{-value} < 0.01$, Stuart-Maxwell test, which indicates the effectiveness of adversarial learning. CD-GAN shows consistent superior result compared with StarGAN, validating that the two-step treatment of generator is beneficial. We also compared our method with 3DSE [41]. Our method achieves comparable result in non-contrast phase (accuracy above 0.95). In addition, CD-GAN achieves better accuracy score 0.92, 0.94, 0.93 on early arterial, venous and delayed phases. CD-GAN corrects misclassification of early arterial scans and venous scan where 3DSE would make mistakes, 0.83 to 0.92 and 0.90 to 0.94 respectively. Table. 4.1 summarizes the accuracy score in the confusion matrix, * indicates the statistical improvement ($p\text{-value} < 0.01$, Stuart-Maxwell test). We also extract the information from Dicom header of each scan and compared methods with the header accuracy. Note some scans are labeled blank or with dose usage instead of phase identities

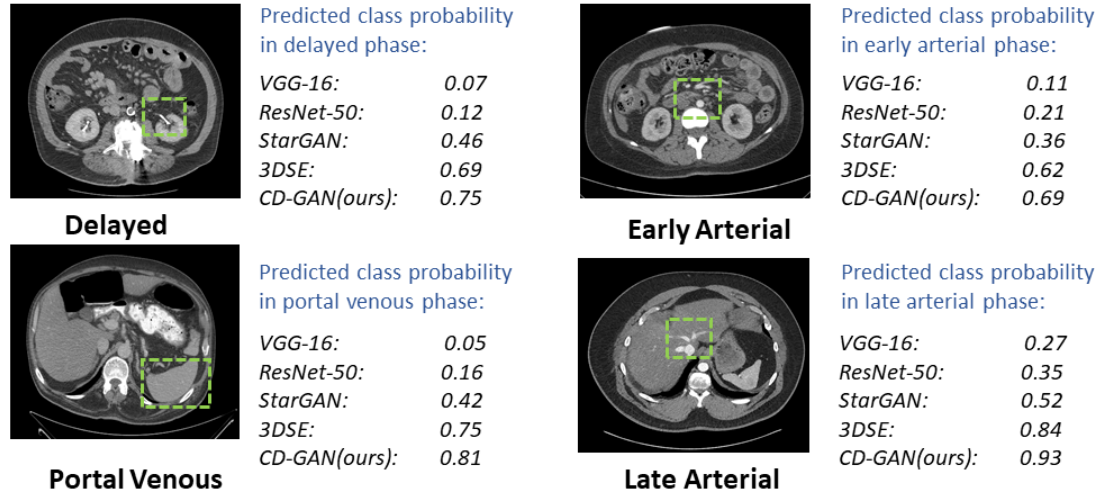


Figure 4.5: Qualitative result of the classification. Four samples are randomly selected from each enhanced phase. Green boxes show critical criteria for identifying enhancement types. Estimated class probabilities for the true label are shown.

in the header, we regard these scans as false ($1(p, p') = 0$) in calculating accuracy scores.

4.3.2 Qualitative Evaluation

Figure. 4.5 shows the qualitative result of the experiment. Four example slices are randomly selected from each enhancement phase. The top left slice shows an image in delayed phase. The green box indicates the illuminated area in urinary system, which is the critical criterion for identifying delayed scans. The bottom left slice is from a patient scanned in portal venous phase, in which liver, spleen and kidneys are slightly bright. This slice is misclassified to late arterial phase by baseline models due to the bright tissues. The top right shows an early arterial slice, the image is misclassified to non-contrast phase by baseline models. The green box shows the only criteria for detecting early arterial phase: the slightly illuminated aorta. The bottom right slice is a late arterial phase image, which portal veins start to be bright (green box). Figure. 4.6 shows the qualitative result on representative error predictions from each type of contrast phase. We observe that non-contrast scans (left) are misclassified into portal venous phase due to higher density tissues of liver or spleen in some subjects. The second image shows a misclassified case in early arterial phase. Due to varying metabolic dynamics, early and late arterial phases are usually difficult to distinguish. In late arterial phase, contrast material has been diffused to lower tissues: portal vein, liver, spleen start

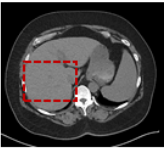
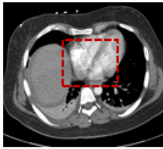
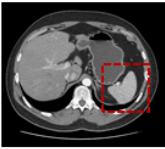
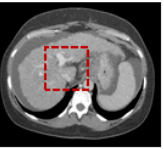
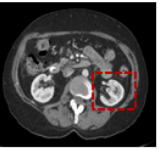
Contrast Phases					
Ground Truth:	Non-contrast	Early Arterial	Late Arterial	Portal Venous	Delayed
Mislabel:	Portal Venous	Late Arterial	Portal Venous	Late Arterial	Portal Venous
<i>Predicted class probability:</i>					
VGG-16:	0.27 0.32	0.32 0.47	0.24 0.46	0.41 0.45	0.36 0.51
ResNet-50:	0.29 0.37	0.34 0.45	0.22 0.45	0.40 0.47	0.32 0.48
StarGAN:	0.31 0.41	0.38 0.51	0.36 0.41	0.46 0.48	0.40 0.45
3DSE:	0.32 0.44	0.42 0.47	0.37 0.42	0.47 0.52	0.37 0.49
CD-GAN(ours):	0.35 0.40	0.46 0.48	0.37 0.39	0.45 0.50	0.42 0.44

Figure 4.6: Representative error prediction in each type of contrast phase. Red boxes show the misleading criteria for mislabels. Predicted class probabilities are shown on both ground truth and mislabels (red) with all methods.

Table 4.1: Quantitative comparison of scan-level performance. We mark * to indicate statistically significant (p-value < 0.01, Stuart-Maxwell test) compared to previous method. Best values are marked bold.

Model	Non contrast	Early Arterial	Late Arterial	Venous	Delayed	Total
Manual HU	0.99	0.95	0.96	0.95	0.99	0.967
Header	0.85	0.31	0.34	0.72	0.84	0.612
VGG-16	0.67	0.42*	0.46*	0.68	0.72	0.590
ResNet-50	0.69	0.46*	0.49*	0.71*	0.74	0.618*
StarGAN	0.76*	0.63*	0.69*	0.75*	0.79*	0.724*
3DSE	0.95*	0.83*	0.89*	0.90*	0.92*	0.898*
CD-GAN	0.95	0.92*	0.91*	0.94*	0.93*	0.930*

to be bright. The effect results in mislabeling with portal venous phase, in which liver and spleen are bright (Figure. 4.6 columns 3 and 4). The fifth column in Figure. 4.6 show a representative error prediction of delayed scans. The bright outer kidney indicates the scan is acquired during venous phase and delayed phase, which raises difficulties in classification. The potential failure predictions on class probabilities resulted in many error classifications. In this experiment, we aimed at acquiring volume-wise phase identifications. Herein, we employed fusion strategy with majority voting. The fusion method effectively addressed many wrong predicted slices, lead to a correct phase corresponding to the CT scan.

4.4 Discussion

In this work, five representative CT contrast phases were classified by comparing different classification models. We revisited the challenge of classifying phase identities using abdomen CT scans. The leading classification model was found to be both generative and discriminative of domain features. In general, GAN architectures showed a better accuracy score with measurement of multi-class statistical test ($p < 0.01$) than conventional models of stacking conv layers. In addition, uncertainties in the classifier may have been reduced when discriminator tries to distinguish real/fake images. In comparison of GAN architectures, we implemented the method of performing multi-domain generative model, StarGAN. The core limitation of StarGAN is it contains a shallower generator to perform image translation with small dimension. Another problem for StarGAN is that the generator fails to capture domain-invariant features (phase identities), which is important in this task. We also compared state-of-the-art model 3DSE, which is a 3D model for detecting phases. 3DSE is originally implemented as 3D model trained on thousands of scans. In our case, CD-GAN is trained on 400 scans, which outperforms 3DSE with consistent improvement (Figure. 4.4). Overall, our method achieved superior performance with three key factors: 1) adversarial learning present effectiveness for helping classifier to capture phase identities, 2) generator perform as both generative and discriminative, and 3) multi-task discriminator design. One of the limitations in our study presents in the training scans since we do not have the complete paired multi-phase scans for all subjects. Capturing the inter-subject variance can be challenging in the experiment, and it is expected to have improved performance if trained with complete paired CT volumes for each patient. In the future, the additional work is needed to continue investing accurate cues for the

task of phase identification.

The experience of training the network is presented. In our experience, GANs converge in the training which stops at about 500,000 iterations (section 2.5), whereas the number for VGG-16 and ResNet-50 is 4,000,000. The largest relative difference between GANs and VGG backbones is the adversarial learning. Those observations are in accordance to findings of previous studies of adversarial examples [150]. Compared with single convolutional networks, adversarial learning shows generally lower sensitivity [150] of variation in dataset. This is probably due to minmax optimization, whose presence leads to boosted performance for both.

In this work, we studied the task of identifying CT contrast phases using abdominal CT scans. We showed that our method achieved consistent superior performance as shown in Figure. 4.5. However, abdominal scans may not cover all the structures that hints the phase stage. For instance, the upper aorta and heart in the lung region is not investigated in the study. It is expected that the cardiac CT imaging can further help determine the phase. The change of brightness of heart and upper aorta relates the contrast timing. In addition, cardiac outputs result in the arrival of contrast medium bolus and affects the timing of enhancements. In the future work, the cardiac imaging studies could be added to the experiment.

We proposed a GAN architecture for contrast phase identity disentangling and provided a robust classifier for body CTs in clinical archives that the classification performance can be boosted with both synthetic images and real images. One major limitation for current image translation or image synthesis is that GANs fail to handle high resolution contexts. Recently, Brock et al proposed BigGAN [151], which studied the generation of high dimensional natural images with huge amount of GPU resources. This study indicated the instabilities of current adversarial learning for image synthesis. In addition, using synthetic images in medical scope should be more careful due to clinical usage. Additional work is needed to investigate possible application of synthetic result. To further improve the detection and classification of contrast phase identities, the Hounsfield unit scale to material density distribution used by reconstruction signals could be adapted to the classification model.

4.5 Conclusion

In summary, the proposed CD-GAN achieved consistent superior performance compared with baseline frameworks. The model developed in this study was used to detect clinically acquired CT images. The configurations are fulfilled by introducing 1) adversarial learning, 2) disentangled representations by generator, 3) multi-task discriminator.

CHAPTER 5

High-resolution 3D Abdominal Segmentation with Random Patch Network Fusion

5.1 Introduction

Computed tomography (CT) of the abdomen is an essential clinical tool in diagnostic investigation and efficient quantitative measurement for internal organs, bones, soft tissue and blood vessels [30]. CT allows for identification of structures in possible abnormalities and tumors. To explore complicated spatial relationship between abdominal organs and tissue structures, multi-organ segmentation on CT scans has been widely studied [19, 49]. Manual annotations are regarded as gold standard [51], but these are time and resource intensive. To reduce the manual efforts, atlas-based methods [50, 52, 54, 152] and deep models [65, 111, 153] have been proposed to achieve quantitative organ segmentation from the clinically acquired CT scans automatically [154]. CNN based models are widely explored for medical image analysis. From the perspective of computation resource-accuracy trade-offs, 2D approaches take separated slices for training result in lacking spatial information, but faster at approximately batch size of 8 and 700 iterations per minute [59]. The tri-planar architecture [155] performs better than 2D with advantage of three views for each voxel at approximately batch size of 3 and 200 iteration per minute. The 3D architecture needs scans to be either 1) patched or 2) down-sampled, it is the slowest way to train at approximately 80 iterations per minute and one patch per iteration. Recently, CNN methods have been explored to 3D segmentation, which perform abdominal segmentation with 3D volumes, like 3D U-Net [57] or V-Net [156]. However, we cannot directly fit the clinically acquired high resolution CT (e.g., 0.8 mm or higher isotropic voxel size) to such networks due to the memory restriction of prevalent GPU. In this context, [157] proposed a network to learn 2D or 3D patches along with volume coordinates for 3D segmentation. One key observation is that the patch-based methods [53, 56, 158, 159, 160] in high-resolution approaches tend to underperform given a lack of broad spatial context. Huo et al. [161] demonstrated that the patch-based method for whole brain segmentation, to deal with the local anatomical variation based on registered atlases.

Unlike brain segmentation, abdomen CT do not have a well-established registration method for

standard space due to larger variations in soft tissues among subjects. Thus, removing spatial context of a high dimensional volume by cropping images leads to a loss of relevant knowledge for abdominal segmentation. A second way for implementing 3D training is to down-sample the image to low resolution volume [57]. However, this approach introduces fuzzy interpolation operations that will break biologic structures in medical images. Holger et al. proposed hierarchical method [66], which introduced a coarse-to-fine strategy that significantly improved the performance of pancreas segmentation. Holger et al. also proposed the multi-scale pyramid network [42] that extend the hierarchical strategy to multi-stage learning. The input images are scaled at different levels, and predictions by last level can be selectively emphasized. However, the performance of scaled images may miss voxels due to inaccurate bounding box predicted by lower level models. Additionally, the output segmentation from upper levels present higher resolution but it still needs to be up-sampled to original space.

Currently, most prevailing deep learning frameworks on medical image segmentation are focused on similar backbones: FCN [153], U-Net [65] and Fast R CNN [162]. In practice, the tri-planar architecture aims to collect combinations of three-view slices for each voxel, and a 3D approach employs a 3D CT scan represented by a sequence of 2D slices. One of the first 3D models was introduced by Urban G et al. [163] to segment brain tumors with varying size. The intuition was followed by multi-scan, multi-path models [164, 165] to capture subsampled features of the image. To exploit 3D context and to cope with limitation of computational resource, researchers investigated hierarchical frameworks. They attempt to extract features at multiple resolution levels. Holger et al. [166] proposed a hierarchical architecture to perceive multi-scale information in pancreas segmentation. Chen et al. [165] aims to simulate human behaviors and generalize RNN to employ 3D context. These approaches provide handling of different field of views at multiple levels, which reduces problems in both spatial context and low-resolution segmentation.

More works have been done on coarse-to-fine methods on abdominal organ segmentation. Li et al. [167] proposed hybrid densely connected UNets, which learns 2D intra-slice features in the first stage then concatenates 3D contexts in the second stage. However, the connected 3D contexts are still down-sampled volumes, which limits in preserving high-resolution details. Zhou et al. [168] developed an FCN based fix-point model to learn both the rough pancreas location and fine segmentation. But it only considered coarse-to-fine regions regardless of overlap regions, which is

not optimized for spatial predictions. Similarly, Roth et al. [66] proposed a hierarchical method, which introduced a coarse-to-fine strategy that significantly improved the performance of pancreas segmentation. However, by constraining rough pancreas locations, the method might be vulnerable to lose information. Roth et al. [42] extended the coarse-to-fine method to multi-scale pyramid networks. The input images are scaled at different levels, and predictions by last level can be selectively emphasized. However, the performance of scaled images may still miss voxels due to inaccurate bounding box predicted by lower-level models. Zhu et al. [63] proposed an effective sliding window approach, performing 3D pancreas segmentation in two stages from both entire CT volume and sub-volumes. In addition, Zhu et al. [63] used the expanding bounding box to improve the robustness for covering target regions. This approach adjusts many outliers in the experiment, but it still may suffer from catastrophic failures in the coarse stage. In summary, current state-of-the-art coarse-to-fine method for abdominal organ segmentation still needs down-sampling in the training and testing and might be vulnerable to failure localization in the coarse stage. To address these issues, we focus on proposing effective patch-based method without isotropic interpolation in the fine stage, while protecting patches from losing target information. Herein, we propose a concise coarse-to-fine framework named random patch network fusion (RPNF), design to alleviate the difficulties for 3D multi-organ segmentation. The method presents two advantages comparing to state-of-the-art methods.

To deal with the anatomical variance from medical images, patches are widely employed to handle the high dimensionality issue [56, 158, 161]. Schelegl et al. [169] proposed a patch-based method on retina images. For medical image segmentation, 3D patch-based methods are used in many applications [170, 171]. In these methods, patches are represented as structural tiling architectures. Each individual region within patch follows fixed pattern over a pre-defined cropping displacement. Ding et al. [172] proposed translational data augmentation, which employed shifts at each sampling point. The resulting performance exploited advantages of successive shifts and yields to final result with concatenation. However, these methods are dependent on manually defined landmarks or extra labels. To break the fixed definition, Cheheb et al. [173, 174] evaluated the benefit of random features for patch-based segmentation, which provided a robustness analysis to patch-based methods. Coupe et al. [175] proposed an ensemble method based on a large number of CNNs processing for different brain areas. The assembleNet [175] introduces sharing of knowledge

among multiple U-Nets and assemblies result with high-resolution predictions by majority voting.

Herein, we propose a concise coarse-to-fine framework named random patch network fusion (RPNF), design to alleviate the difficulties for 3D multi-organ segmentation. The method presents two advantages comparing to state-of-the-art methods. 1) it enables segmentation in original CT resolution without image scaling in the input and output. 2) it performs robustness to save the catastrophic failures from the coarse stage. The method enables segmentation in original CT resolution by a two-stage cascade design. The proposed strategy is built on the concept that the performance of a higher resolution level in hierarchical model is indicative of the low-resolution level in hierarchy. To validate the proposed strategy, experiments on baselines methods are performed, including low-resolution [57], high-resolution [158] and multi-scale pyramid models [176]. For the family of patch-based method, we evaluate different strategies including structural tiling, random shifting and combined approaches. We perform sensitivity analysis in terms of patch numbers, as well as the ablation studies on behalf of averaged coverages per voxel and the effect of variant patch size. We present our study on the dataset from “Multi-Atlas Labeling Beyond the Cranial Vault” (BTCV) Challenge of MICCAI 2015. For external validation, we evaluate our method on two cohorts that were excluded from training, the ImageVU pancreas with 40 subjects and HEM1538 with 82 subjects. To summarize, the contributions of this work are:

1. We proposed a new coarse-to-fine framework termed ‘random patch network fusion’ by introducing randomly localized patches between first and second stage.
2. We show that our proposed method can be implemented to predict original space segmentation in second level model.
3. We provide large-scale validations on analyzing patch-based strategies and comparing them with our method, supporting that patch-based method plus random shifting could boost 3D segmentation performance.

5.2 Theory

The proposed method for abdominal segmentation consists of three main components: (1) a 3D multi-organ U-Net that produces coarse, preliminary segmentation, (2) a random patch sampling process which imposes constraints of the field of view, and (3) a second stage model followed by

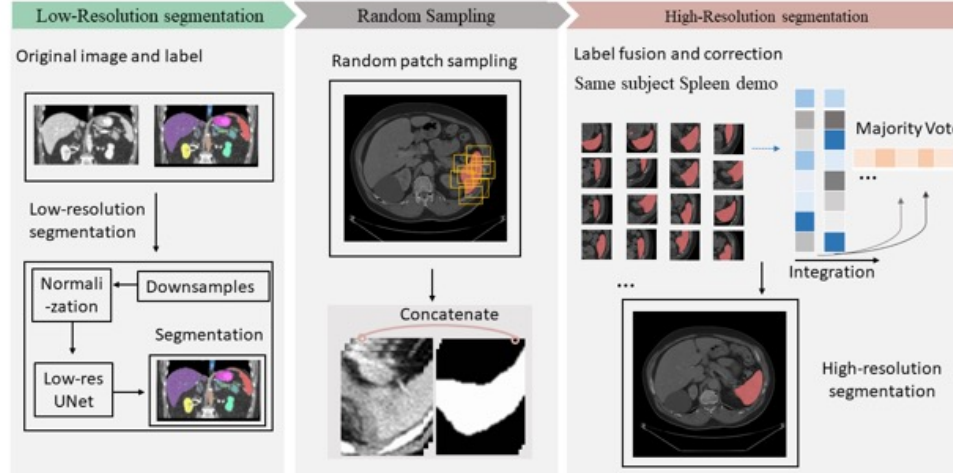


Figure 5.1: Method framework. Given a CT scan with at high resolution of $0.8 \times 0.8 \times 2\text{mm}$, a low-res section (left panel) is trained with multi-channel segmentation. The low-res part contains down-sampling and normalization in order to preserve the complete spatial information. After the coarse segmentation are acquired from low-res UNet, we interpolate the mask to match the image's original resolution. Next, random patch sampling (mid panel) is employed to collect patches, and patches are concatenated with corresponding coarse segmentation masks. Finally, we trained a patch-based high-res (right panel) segmentation model, the high-dimensional probability maps are acquired from integration of all patches on field of views. Majority vote is used to merge estimates into a final segmentation.

statistical fusion to achieve final segmentation (Figure. 5.1). The approach combines convolutional neural networks, hierarchical models and statistical fusion. For training the subject image i , given Hounsfield Units of voxels, the goal of the random patch network fusion algorithm is to estimate the segmentation S using observed labels s' from voters V_i . Consider the framework as a hierarchical model with two stages l and h . At each level, let $S_m = (S_l, S_h)$ be the mapping vector that corresponds to labels at each level of segmentation. Let $s \in L$, $Y_m = (Y_l, Y_h)$ be the collection of ground truth at the m level of hierarchy. The entire problem definition of our goal is to estimate the segmentation such that:

$$f(V_{ij} = s' | Y_i = (s, S, \theta)) \quad (5.1)$$

where voters V for each voxel j observes label s' given the ground truth Y , hierarchical model S , the parameters of each model θ .

5.2.1 Stage 1: Preliminary segmentation

The labeled data i represents the original resolution CT scan. x is the down-sampled volume using tri-linear interpolation. Consider the segmentation network parameterized by θ . Low-resolution training aims to:

$$\operatorname{argmax}_{(\theta, Y_i)} L_{D_i}(\theta) \quad (5.2)$$

where $L_{D_i}(\theta)$ is the Multi-Sourced Dice Loss (MSDL) [4]. MSDL was proposed as a way of evaluating datasets with varying labels with a single score by extending the Dice loss to adapt unbalanced multi-organ segmentation:

$$L_{D_i} = -\frac{2}{A} \frac{\sum_{a=0}^A w \sum_{i=1}^M \sum_{j=1}^N Y_{ij} P_{ij} + \epsilon}{\sum_{a=0}^A w \sum_{i=1}^M \sum_{j=1}^N Y_{ij}^2 + \sum_{a=0}^A w \sum_{i=1}^M \sum_{j=1}^N P_{ij}^2 + \epsilon} \quad (5.3)$$

where A denotes the number of anatomies and w represents the variance to different label set properties in given image dimension of M and N . Y is the voxel value and P are the predicted probability maps. A small number, ϵ , was used in computing the prediction and voxel value correlation to prevent discontinuities. MSDL was iteratively optimized, and P_{ij} was computed by the softmax of the probability of voxel j in image i to anatomy a .

5.2.2 Stage 2: Random patch sampling

We proposed an approach that is inspired by hierarchical algorithms [171] and random sampling [173]. We randomly select predicted voxels in the coarse segmentation mask according to the distribution. Using the selected voxel as indices' center, we place a bounding box as the local field of view. In order to introduce randomness, we also add a random shift to all axes' direction by the distribution. The distance of shifting is given by Gaussian random number generator, the mean and variance of the norm is determined by the mean distance of centers indices (e.g. the spleen patches have the mean distances of 4.2 and variance of 2.3 in x axes direction, the shifting distance to x direction is generated by the Gaussian random number generator). Patches are cropped according to bounding boxes as second stage model inputs as shown in the middle panel of Figure. 5.1. The strategy crops CT scan at original resolution without re-sampling, and it builds the hierarchy of non-linear features from random patches regardless of 3D contexts. The method employs detail context at original resolution and incorporates advantages of data augmentation with shifting.

5.2.3 Stage 3: Label fusion

After separating full spatial context to k randomly selected subspaces, patches will overlap with each other. The overlapped region could provide more than one segmentation label for a voxel. Herein, except placing patches back to original coordinate space, it is required to summarize a single label given a vector of class labels from n candidates. In this work, we implement label fusion with majority vote algorithm, which fuses n segmentation from network predictions to a single label. The final segmentation label for voxel j in image i is acquired by:

$$S_{ij} = \operatorname{argmax} \frac{1}{n} \sum_{m=1}^n p(a|s'_m, j) \quad (5.4)$$

where $p(a|s'_m, j) = 1$ if s'_m equals to anatomy class a and 0 otherwise. We ignore the voters outside the image space, related values are excluded in the label fusion. For voxels with equal number of voters, we label the voxel randomly to either be target or background. uncertainty.

5.3 Methods

The 3D abdominal segmentation task involves segmentation of 12 abdomen structures with highly deformable volume and shape. Anatomies present high class-imbalance, which involve large organs (spleen, liver, stomach and kidneys), vessels (aorta, portal and splenic vein, and inferior vena cava (IVC)), and small anatomies (esophagus, gallbladder, pancreas and adrenal glands). The details of each dataset are provided below.

BTCV dataset: We perform de-identified data acquired from the Vanderbilt University Medical Center (VUMC) under IRB approval. We retrieved 100 subjects with 12 labeled anatomies, labels are annotated by experts. We integrate all 100 subjects in this study, the in-plane pixel dimension of each scan varies from 0.4 to 1.2 mm. Each volume is preprocessed by excluding outlier intensities beyond -1000 and 1000 HU. The slice thickness ranges from 1 to 6 mm. Each CT scan consists 80 to 225 slices of 512 x 512 pixels. 100 CT scans are independent from subjects and with contrast enhancement in portal venous phase. Part of the dataset is released in the MICCAI 2015 Multi-Atlas Labeling Challenge, which contains 30 scans with 3779 axial slices. The 12 organs were outlined manually by interpreters under supervision of clinical radiologists (MD) from Vanderbilt University Medical Center (> 10 years of experience in abdomen radiology). For each organ, the interpreter

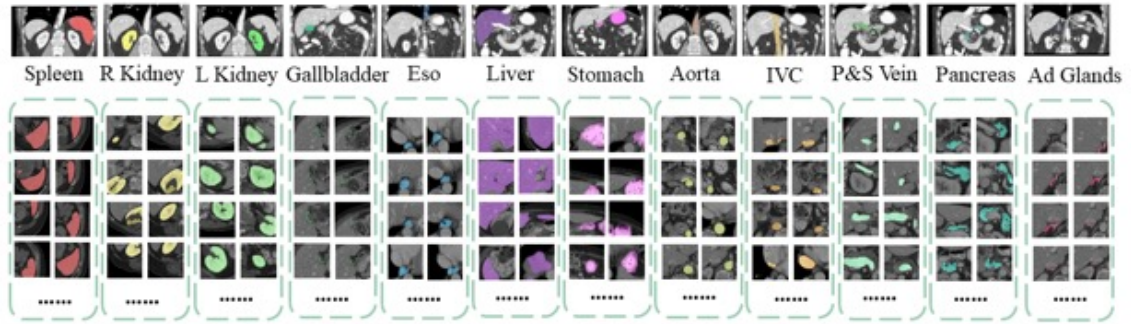


Figure 5.2: Representative random patches for 12 abdominal organs of a single subject. The patch size is $128 \times 128 \times 48$ and 8 samples are shown for each anatomy. Patch size defines the volume of field of view corresponding to organs. Large organs like spleen, liver and stomach cannot be covered until a number of patches are sampled, the patch of $128 \times 128 \times 48$ covers most regions of mid-sized organ (kidney, pancreas, and portal & splenic vein), while small anatomies (adrenal glands, gallbladder, and vessels) can be covered by single patch with above size. The patch size effect is explored in an ablation study.

was instructed to verify the segmentation slice-by-slice in all axial, sagittal, and coronal views. To avoid inter-rater variability and perform reproducibility, we have independent observers perform manual segmentation on the same dataset.

HEM1538 dataset: We retrieved 82 splenomegaly subjects substantially acquired with clinically trials. 117 splenomegaly CT scans are included and used as external validation [141]. Splenomegaly indicates the enlargement of spleen with different levels of red blood cell destruction and inflation. These scans have large variance of spleen shape, which size varies from 143 cubic centimeter (cc) to 3045 cc. Each CT volume consists of 60 to 200 slices of 512×512 pixels, with resolution of (0.59×0.59 mm to 0.98×0.98 mm). The slice thickness ranged from 1mm to 2mm. For each case, the spleen is manually annotated and reviewed by a radiologist.

ImageVU pancreas: A total of 40 subjects were selected and retrieved from Vanderbilt University Medical Center (VUMC). The dataset is collected from a group of 40 outliers out of 598 retrieved subjects. These outlier-guided subjects were a collection of studies that evaluated with benefits for rare/in-frequent population. The pancreas was manually traced for each subject under a soft tissue window. *Patches for organs:* To illustrate the random patch definition, we show the sampled patch field of views in Figure. 5.2, which are acquired from BTCV dataset, 12 annotated anatomical structures are shown.

5.3.1 Preprocessing and body part regression

We processed CT scan with soft tissue window with range of -175 to 250 HU. Intensities were normalized to (0,1). Clinically acquired CT scans can exhibit large variance in volume size, we adopt a critical pre-processing with body part regression [31, 114]. The body part regression helps to remove slices on inconsistency volumes, and to localize anatomical regions automatically. We used the pre-trained model from unsupervised regression network to navigate slices in abdomen region (scalar reference index ranges from -6 to 5).

5.3.2 Baseline architectures

We compared our random patch network fusion framework with a series of state-of-the-art approaches, including 1) low-resolution model on down-sampling images to fit maximum GPU memory, 2) high-resolution architecture with complete tiling patches, and 3) hierarchy with two-level pyramids.

Low-resolution architecture: The 3D U-Net is trained on images with finest resolution to house the maximum GPU memory. Each scan is down-sampled from (512, 512) to (168, 168) and normalized to consistent voxel resolution of (2 x 2 x 6). The output and ground truth labels are compared using MSDL. We ignored the background loss in order to increase weights for anatomies. The final segmentation maps are up-sampled to original space with nearest interpolation [177] in order to spatially align with CT resolution. This approach is trained end-to-end, and the resulting segmentation is summarized in Figure. 5.3 and Figure. 5.4. Qualitative results are shown in Figure. 5.5 and Figure. 5.6. The low-resolution framework incorporated down-sampled volume which lacks detail structures of anatomies, but it preserves complete spatial context in CT scan.

High-resolution architecture: The image is normalized to 1 mm isotropic resolution with dimension of 512 x 512. Since the high-resolution volume cannot be fed into GPU given structure of 3D U-Net, we employed k patches tile to cover full CT space. The patch number k for each image is based on equal distribution that continually tiles in x, y, and z axes. Each image is split to 32 patches along with the dimension, each patch covers a subspace. To maximize the usage of GPU memory, we use patch volume with (168 x 168 x 64) voxels. The subspace can be presented

by coordinate (x, y, z) and patch size (d_x, d_y, d_z) .

$$\phi_k = [x_k : x_k + d_x, y_k : y_k + d_y, z_k : z_k + d_z] \quad (5.5)$$

Patches are extracted without overlaps, each patch is padded to fixed size once it exceeded the volume dimension. For 3D U-Net, we adjust the decoder section upon original 3D U-Net implementation to be compatible with 12 labels prediction. 12 output channels are employed in the de-convolutional layers in the model. We also presented the overlapped patch strategies analyzed in Figure. 5.7. The half-overlapped patches covered half volume of subspaces, one-third overlapped patches cover one-third volume of subspace, etc. The high-resolution method is evaluated end-to-end, and final segmentation masks are acquired by tiling ordered patches.

Hierarchy with multi-scale pyramid network: To effectively segment an image at higher resolution, we compare our method with the multi-scale auto-context pyramid approach. The method both captured spatial information at lower resolution down-sampled images while learned accurate segmentation from higher resolution in multiple levels. $F = (f_m(X_m, \theta_m)), m = 1, \dots, M$, with m the order of levels in the approach. X_m is the subspace at level m . In the first level, the 3D U-Net is trained the same as low-res network, which employed lowest resolution to fit largest amount of spatial information. In the next levels, it uses the predicted segmentation masks as an input channel to the next network. The succeeded input volume is cropped according to bounding box define by predicted segmentation map in level $m - 1$. And down-sampled by a factor of $d_m = d_{m1}/2$. The previous level’s segmentation is up-sampled by 2 in order to align with higher resolution levels. 3D U-Net at each level is optimized using Dice loss as the same wit low-resolution training. The predicted segmentation masks and cropped images are concatenated as the next level input. The final segmentation masks are acquired by interpolating the last level prediction and match the cropped images to original coordinates. In our implementation, we trained the pyramid models with two levels.

5.3.3 Implementation Details

We adopt 3D U-Net as the segmentation model, which contains encoder and decoder paths with four levels resolution. It employs deconvolution to up-sample the lower level feature maps to the higher

space of images. This process enables the efficient denser pixel-to-pixel mappings. Each level in the encoder consists two $3\times 3\times 3$ convolutional layers, followed by rectified linear units (ReLU) and a max pooling of $2\times 2\times 2$ and strides of 2. In the decoder, the transpose convolutions of $2\times 2\times 2$ and strides of 2 are used. And followed by two $3\times 3\times 3$ convolutions, followed with ReLU. 3D U-Net employs skip connectors from layers of same level in the decoder to provide higher-resolution features to the decoder part. The last layer is a $1\times 1\times 1$ convolution that set the number of output channels to the number of class labels. We used Multi-sourced Dice Loss and Dice Loss for multi-organ segmentation and single class segmentation respectively.

The baseline low-resolution multi-organ segmentation uses the largest volume size of $168 \times 168 \times 64$ in order to fit maximum memory of a normal 12GB GPU under architecture of 3D U-Net. The volume size is also employed in baseline hierarchical method for training the first level model. For patch-based segmentation, we firstly chose the medium size of $(128,128,48)$ for experiments, the effect of different size of patch is evaluated in ablation study, presented in Figure. 5.8 and Figure. 5.9.

To fairly compare methods, the same 3D U-Nets is used with same hyper-parameters except input dimension and channels. We use batch size of 1 for all implementations. We used Instance Normalization, which is agnostic to batch size. We adopted ADAM algorithm with SGD, momentum=0.9. The learning rates is set to 0.001 and it reduced by a factor of 10 every 10 epochs after 50th epoch. Implementations are performed using NVIDIA Titan X GPU 12G memory and CUDA 9.0. Training, validation and testing are executed on a Linux workstation with Intel Xeon CPU, 32GB of RAM. The code of all experiments including baseline methods are implemented in python 3.6 with anaconda3. Networks and frameworks are implemented in Pytorch 1.0.

5.3.4 Experimental Design

We conducted experiments on three perspective of analysis to evaluate the effectiveness of different approaches. First, we compared state-of-the-art methods with RPNF on multi-organ segmentation to provide effectiveness. Then, in order to prove robustness and sensitivity, we did three ablation studies to validate the effect of 1) patch-based strategies, 2) number of random parches per scan and 3) patch size. Last, we tested the trained model on two external datasets to provide stability of the RPFN. All segmentation comparisons are assessed the average DSC score across 12 non-

background labels. The claim of statistical significance is evaluated by paired t-test ($p < 0.01$).

5.3.4.1 Random Patch Network Fusion

To perform best effectiveness of the proposed method, we implemented experiments with maximum number (50) of patches for evaluating performance of baselines and RPNF. We implemented experiments with 5 fold cross validation on BTCV dataset. To perform standard five-fold cross validation, we split 100 scans into five complementary folds, each of which contains 20 cases. For each fold evaluation, we use 4 folds as training and testing on the remaining cases.

We compared RPNF with three baseline architectures (low-resolution, high-resolution and hierarchy) with same dataset and parameters on task of multi-organ segmentation. Briefly, we first trained the low-resolution approach, which has been shown its capability on 3D multi-organ segmentation with full spatial contexts. Second, we trained the 3D networks with structural tiles without overlapping to evaluate the high-resolution method. In this setting, local patches are cropped by pre-defined coordinates, patch size remains the same as low-resolution training (168 x 168 x 64). To evaluate the coarse-to-fine methods, we trained the two-level pyramid networks as the third baseline. In the first level, the low-dimensional representation is used for computational efficiency. Then, the input volume is cropped according to bounding box define by predicted segmentation map in the first level. The volume is down-sampled by a factor of $d_2 = d_1/2$, which means pyramid networks use higher resolution patches in the second level hierarchy. In this experiment, we evaluate the proposed method on all of 12 abdominal anatomies. We aim to show the method can be applied to multiple organs with variant sizes (large organ such as liver, and small anatomy such as adrenal gland). Thus, we claim the effectiveness of RPNF on the representative multi-organ dataset.

5.3.4.2 Ablation Study

In this section, we evaluated the effect of three key factors that influence the performance. To simplify the evaluation on patch strategies and numbers, we conducted experiments on spleen segmentation as the representative task of abdomen segmentation. On evaluation of patch size, we performed the experiments on three representative organs (liver, spleen and pancreas). We performed experiments on the same BTCV dataset with 80 training scans, the withheld 20 cases are split with 10 for validation and 10 for testing.

Effect of patch-based strategies: For comparing patch-based strategies, we implemented methods of structural tiling, structural plus randomness and pure randomness. Briefly, we first employed complete structural tiling. Similar to high-resolution training, we cropped the image with fixed coordinates. In the second strategy, we start from the structural bounding boxes in the first method, then randomly shift each box in three directions (x, y and z). Last, we perform pure random selection of patches instead of pre-defined bounding boxes. To perform a fair comparison, the same 3D segmentation network with the same parameters are used in experiments. To be specific, the patch size = 128 x 128 x 48, batch size = 1, optimizer = “Adam”, and learning rate = 0.001. The preprocessing remains the same for different strategies.

Effect of average number of coverages per voxel: To further evaluate patch-based methods, we conducted a large-scale of experiments on number of coverages per voxel. We designed experiments on use of average number of coverages per voxel from 1 to 50. The structural tiling method is implemented by increasing overlapped region with 1/2, 1/3, 1/4, ... 1/50. The structural plus randomness is performed by shifting overlapped patches randomly. Last, the evaluation on the pure randomness strategy is achieved by increasing number of random patches until reach the same average coverage per voxel as other experiments. We showed the analysis in Figure. 5.7.

Effect of patch size: While patch-based studies achieved promising results, there are rare work focused on effect of patch size. Following the current prevailing GPU memory, previous studies implemented patch with dimension for maximizing usage of memory (for example: 128x128x96 volume typically occupies 12 GB with 3D U-Net). To better understand the effect of patch size, we evaluated 3D segmentation on three representative abdominal organs, spleen, liver and pancreas. We first employed three different sizes for evaluating dimension of x-y axes. Volume size varies from 64x64x48, 96x96x48 to 128x128x48. Then we performed training on different volume length (128x128x36, 128x128x48 and 128x128x64). Except patch size, other settings remain the same as random patch network fusion.

5.3.5 Validation on External Datasets

To show the stability of the proposed method, we validate the trained model on two independent cohorts (HEM1538 and ImageVU pancreas). We implemented the model of RPNF and baseline approaches on all subjects in these two unseen datasets. For evaluation of HEM1538, we aim to

present the stability of our model, which trained on normal spleen and test on splenomegaly cases. For ImageVU pancreas, we compare the performance on outlier-guided study with academic controlled dataset. The same preprocessing and patch selections are used in the validation correspond to each model. Except the proposed PRNF method, the output segmentation volumes from other models were resampled back to the original image space.

5.3.6 Evaluation Metrics

We used the Dice similarity coefficients (DSC) as the measurement for our method and baseline approaches,

$$DSC = \frac{2|A \cap M|}{(|A| + |M|)} = \frac{2|TP|}{(2|TP| + |FP| + |FN|)} \quad (5.6)$$

where TP is true positive, FP is false positive, FN is false a negative. The statistical measurement between methods were evaluated by paired t-test and the difference was significant when $p < 0.05$.

5.4 Results

5.4.1 Random Patch Network Fusion

In Figure. 5.3, the quantitative boxplot shows the proposed random patch network fusion method with 12 anatomies' labels achieved superior performance compared with baseline methods on metric of DSC scores. Table. 5.1 reports mean DSC scores and standard deviation. As shown in Table. 5.1, our proposed framework achieves state-of-the-art method "Hierarchy" [62] by a large margin. For large organs, our RPFN achieves 0.963 against 0.942 (spleen), 0.965 against 0.955 (liver), 0.856 against 0.826 (stomach), which are around advancement of 1.5%. In comparison of middle sized organ, our methods achieve 0.931 vs 0.881 (right kidney), 0.945 vs 0.887 (left kidney), 0.788 vs 0.758 (esophagus), 0.923 vs 0.903 (aorta), 0.853 vs 0.833 (IVC), 0.761 vs 0.721 (pancreas), which are around 3% advancement. Regarding of small tissues, our method achieves 0.826 vs 0.501 (gallbladder), 0.728 vs 0.698 (portal and splenic vein), 0.736 vs 0.690 (adrenal glands), which increases by a large margin.

Figure. 5.3 indicates our method achieved significant improvement with paired t-test of p-value $< 4 \times 10^{-4}$, compared with performance of the state-of-the-art two-level hierarchy. In Figure. 5.3, the DSC scores of high-resolution methods are the lowest since local patches result in holistic information. Unlike patch-based method in brains [55], abdomen CTs do not have registration step

Table 5.1: Mean DSC and variance of 12 abdominal organs compared with our method and three baseline approaches on BTCV miccai2015 challenge testing cohort. Our method presented significant improvement compared to Hierarchical method. (p-value < 0.01 with paired t-test). Note: Bold values indicates best mean DSC of each organ.

Organ	High-resolution	Low-resolution	Hierarchy	RPFN (ours)
1.spleen	0.8732 ± 0.0316	0.9382 ± 0.0244	0.9422 ± 0.0048	0.9635 ± 0.0050
2.right kidney	0.7675 ± 0.0617	0.8996 ± 0.0180	0.8810 ± 0.0233	0.9310 ± 0.0231
3.left kidney	0.7579 ± 0.0812	0.8893 ± 0.0141	0.8872 ± 0.0435	0.9453 ± 0.0210
4.gallbladder	0.3565 ± 0.0896	0.5394 ± 0.0896	0.5014 ± 0.1178	0.8263 ± 0.0348
5.esophagus	0.6079 ± 0.0139	0.7481 ± 0.0140	0.7582 ± 0.0120	0.7881 ± 0.0120
6.liver	0.9321 ± 0.0039	0.9558 ± 0.0031	0.9557 ± 0.0072	0.9656 ± 0.0670
7.stomach	0.6641 ± 0.0247	0.8298 ± 0.0158	0.8267 ± 0.0118	0.8567 ± 0.0118
8.aorta	0.8540 ± 0.0024	0.9063 ± 0.0241	0.9032 ± 0.0326	0.9232 ± 0.0321
9.IVC	0.7528 ± 0.0068	0.8285 ± 0.0068	0.8328 ± 0.0192	0.8528 ± 0.0186
10.P&S veins	0.5778 ± 0.0123	0.6817 ± 0.0145	0.6979 ± 0.0672	0.7279 ± 0.0670
11.pancreas	0.5581 ± 0.0265	0.6765 ± 0.0265	0.7209 ± 0.0205	0.7608 ± 0.0535
12.Ad gland	0.3956 ± 0.0295	0.6321 ± 0.0295	0.6897 ± 0.0642	0.7356 ± 0.0367

Table 5.2: Segmentation performance of models trained on BTCV dataset in Mean DSC and variance, tested on HEM1538 and ImageVU pancreas, the proposed method is compared with baselines (p-value < 0.01 with paired t-test between random patch network fusion and two-level hierarchy).

	HEM1538 (spleen)		ImageVU (pancreas)	
	Vol DSC	Std	Vol DSC	Std
High-resolution	0.9134	0.0283	0.5046	0.0711
Low-resolution	0.9268	0.0211	0.5537	0.0513
Two-level Hierarchy	0.9493	0.0189	0.5782	0.0548
Random Patch network fusion	0.9672	0.0143	0.6019	0.0423

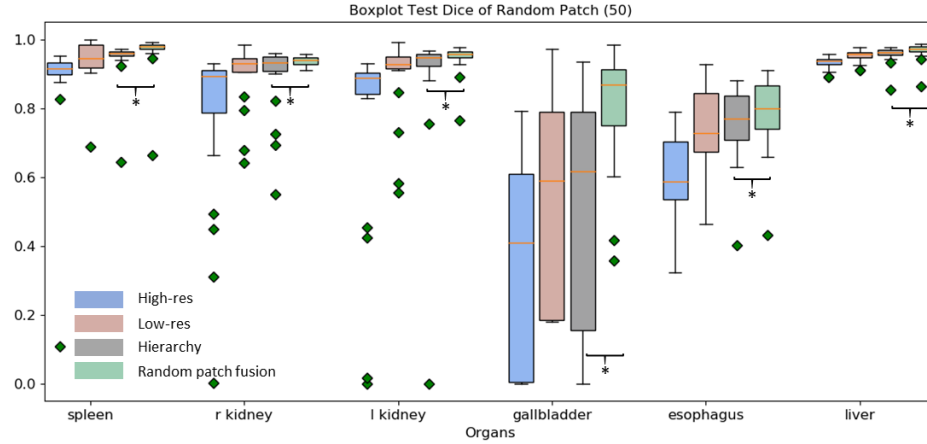


Figure 5.3: Quantitative results from the testing cohort: spleen to liver (50 patches used). We compare our random patch network fusion method with three baseline approaches (high-res, low-res and hierarchical framework). The high-res method presents result with large variance and outliers in boxplot due to limited field of view in each patch. The low-resolution segmentation performs better than high-res method in mean DSC, which indicates complete spatial information is essential in abdominal organ segmentation. The hierarchical approach increases training resolution in the second step and achieved higher DSC. Hierarchical method’s performance is limited when bounding box is inaccurate from previous levels. Our method achieves overall highest result compared to hierarchical method with significant improvement, “*” indicates statistically significant ($p < 0.01$ from paired t-test). The random patch fusion framework employs advantages from both low-res and high-res settings, and it achieves segmentation without resample postprocessing. In boxplot, small anatomies (gallbladder, esophagus) present higher improvements than large organs (spleen, kidneys and liver), which presents higher median DSC, smaller variance and fewer outliers.

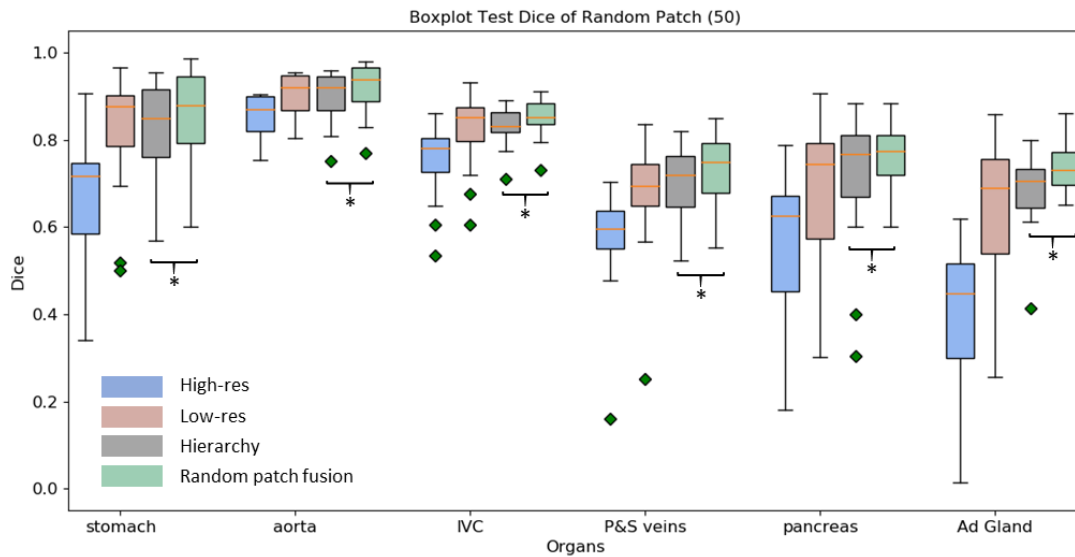


Figure 5.4: Quantitative result for the testing cohort: stomach to adrenal glands (50 patches used). “*” indicates our method outperforms hierarchical method by statistically significant improvement ($p < 0.01$ from paired t-test).

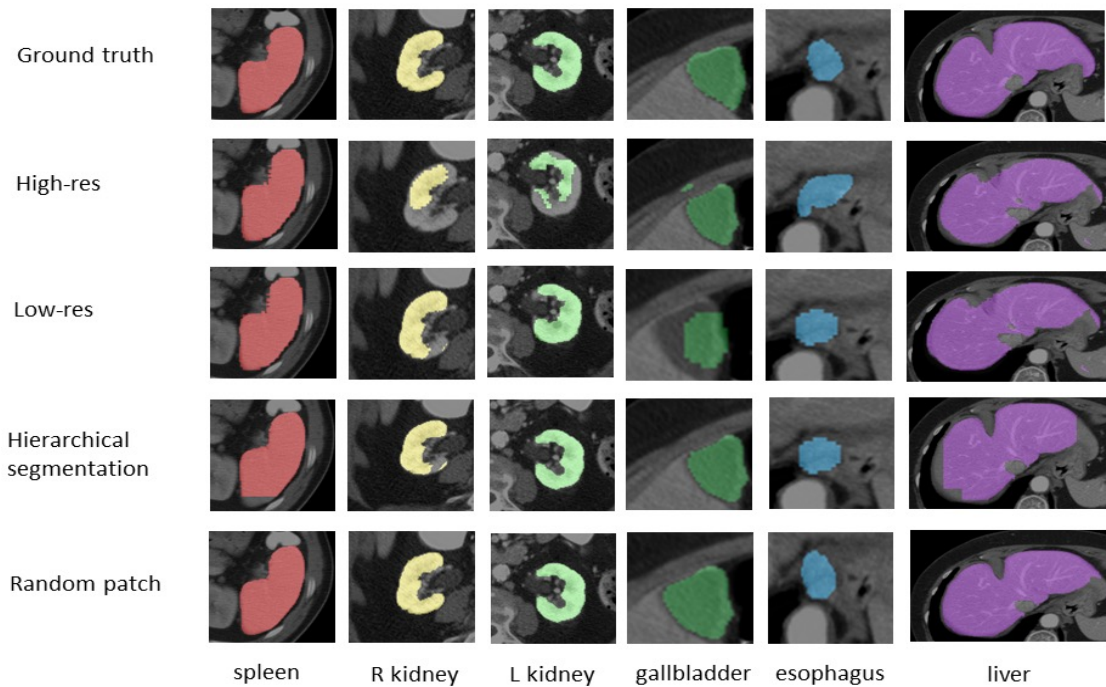


Figure 5.5: Same subject qualitative result of our method compared to baseline approaches (spleen to liver). Second and third row presents direct high-resolution and low-resolution segmentation, mis-predictions are shown due to limited field of view, and resampling respectively. The hierarchical method presents smoother boundaries but suffers from truncation due inaccurate bounding box from first step. Our random patch fusion method presents complete segmentation masks with smoother boundaries among structures.

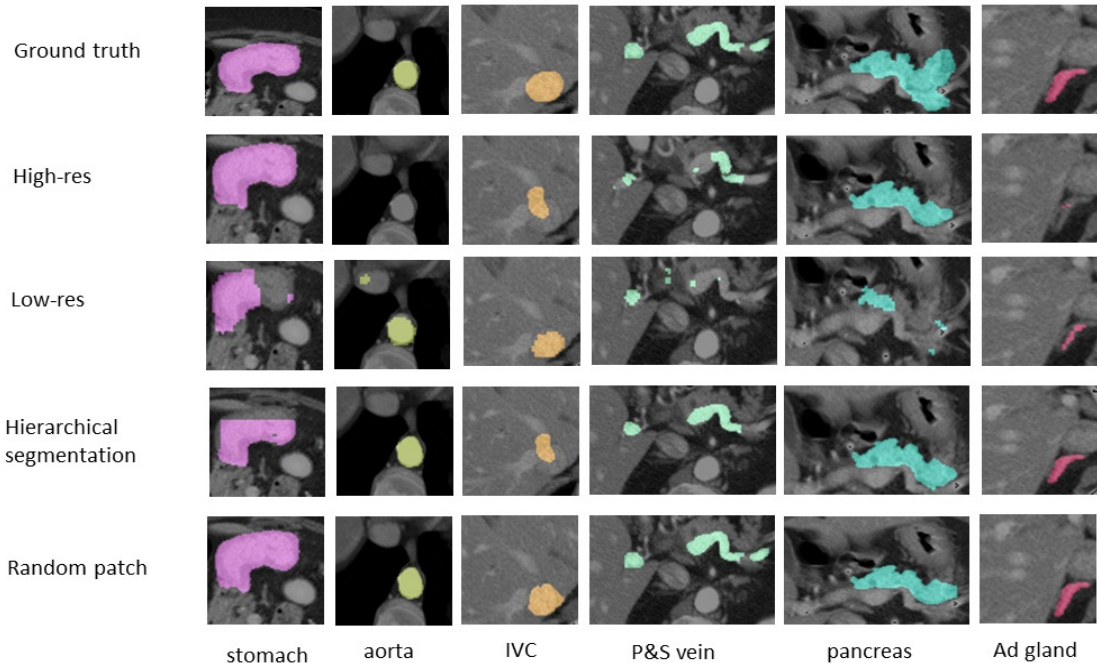


Figure 5.6: The same subject qualitative result of our method compared to baseline approaches from stomach to adrenal glands.

Table 5.3: Comparison of coarse-to-fine methods between our proposed approach and state-of-the-art methods. The evaluation is conducted on BTCV testing dataset in terms of mean DSC.

Methods	spleen	R Kid	L Kid	Gall	Eso	liver	Sto	aorta	IVC	Vein	Pan	AG	All
Roth <i>et al.</i> (Roth et al.2017)	.926	.884	.889	.531	.724	.953	.819	.884	.823	.687	.720	.664	.792
Zhou <i>et al.</i> (Zhou et al. 2018)	.941	.918	.932	.603	.753	.964	.842	.907	.820	.689	.722	.675	.814
Li <i>et al.</i> (Li et al. 2018)	.957	.917	.924	.636	.760	.963	.840	.901	.821	.697	.726	.669	.817
Zhu <i>et al.</i> (Zhu et al. 2018)	.961	.928	.932	.693	.772	.964	.849	.913	.837	.698	.762	.684	.833
Ours	.963	.931	.945	.826	.788	.966	.857	.923	.853	.728	.760	.736	.856

R Kid: right kidney, L Kid: left kidney, Gall: gallbladder, Eso: esophagus, Sto: stomach, IVC: inferior vena cava, Vein: portal and splenic veins, Pan: pancreas, AG: adrenal gland.

to alleviate the bias and variance in patients, which indicates intensities in patches are not scaled and normalized. Herein, we observed that soft structures such as stomach and pancreas show large std (DSC score) in Figure. 5.3 and Figure. 5.4. A similar result happens in large structure, liver and spleen, since the segmented tiled patches contain outliers. On performance of low-resolution model, we see an improvement for all structures compare to high-resolution model, which indicates that spatial contexts are essential for 3D segmentation. As full context provides complete shape and background knowledge to training model, the low-resolution model shows smaller standard deviation in Table. 5.1. The limitation of the low-resolution method comes from the tri-linear and nearest interpolation during downsample-upsample steps. Small structures, gallbladder, adrenal glands are

Table 5.4: Fine stage performance comparison with state-of-the-art methods on patch selection strategies using same backbone network (3D UNet). The evaluation is performed on BTCV testing data on 12 abdominal organs in terms of mean and variance.

Methods	Mean Surface	Average Dice	Hausdorff Distance
<i>Local patches (tiling no overlap)</i>	6.6129± 3.1458	0.6748± 0.0670	52.1484± 31.9348
<i>Local patches (tiling 1/2 overlap)</i>	5.5195± 3.0981	0.7075± 0.0664	47.2357± 26.7541
<i>Kim et al. (kim et al., 2020) (uniform crop)</i>	5.4912± 3.0385	0.7493± 0.0659	45.0924± 27.1705
Roth et al. (Roth et al., 2018b) (fine-scaled)	4.6011± 2.5651	0.7991± 0.0623	38.5917± 20.6583
Zhu et al. (Zhu et al., 2018) (sliding window)	1.8143± 1.0359	0.8297± 0.0617	24.5591± 17.4515
Ours (random patch)	1.4237± 0.5916	0.8564± 0.0608	18.9862± 12.4169

with limited number of voxels in low-resolution volume, the predicted segmentation will not memorize the shape structure after up-sampling with nearest interpolation. In hierarchy approach, we implemented multi-scale pyramid network with two levels, the result present in grey boxes in Figure. 5.3 and Figure. 5.4. The hierarchical approach shows general better DSC scores than low-res model by incorporating spatial context in first level and higher resolution context in second level. The final segmentation result relies on the bounding box predicted by outputs in previous level. We observe that cases may miss part of structure due to cropping with inaccurate bounding box, as presented in qualitative result (Figure. 5.5). The uncertainty of boundaries results in amounts of outliers especially in segmenting soft structures (such as stomach and pancreas). Herein, the boost in DSC score for these structures are marginal for the hierarchical approach. The random patch network fusion presents overall higher DSC scores on all structures. We see a high improvement of 30% DSC score for gallbladder. Adrenal glands also present large improvements, these small structures benefit greatly since a single random patch will cover entire structure, and the random patch works as data augmentation scheme while benefits with assembled result with label fusion. The random patch fusion network utilized advantages from all the baseline approaches, 1) complete spatial context in low-resolution model, 2) detailed feature of structures in high-resolution model and 3) coarse attention mechanism provided by multi-scale architecture. Additionally, the second step of our method predicts masks in original CT space, which indicates no re-sampling step needed for final segmentation result.

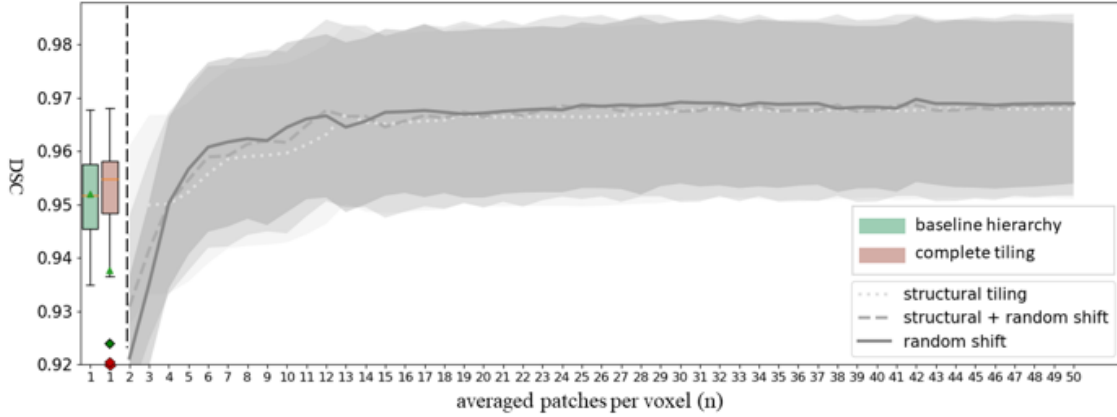


Figure 5.7: Boxplot and uncertainty curves on patch strategies. The boxplots on the left presents the DSC coefficients on testing scans of baseline hierarchy method compared to the complete tiling. Complete tiling shows less variance than baseline hierarchy method. Red diamonds present outliers in complete tiling, and baseline hierarchy shows better DSC (green triangles) than complete tiling. Uncertainty plots show means/standard deviations comparison of structural tiling, structural tiling plus random shift and only random shift methods along with averaged patches per voxel. This presents DSC of each experiment with averaged covered voxels from 2 to 50.

5.4.2 Ablation Study

5.4.2.1 Effect of Patch-Based Strategies

Hierarchy vs complete tiling: We present the results of 3D spleen segmentation by two patch-based baselines in left panel of Figure. 5.6, which is the two-level hierarchy method and complete tiling. The two-level hierarchy approach used the same patch configuration in the second step as complete tiling. Patch size of (128,128,48) is used for both experiments. The implementations are conducted with averaged patches per voxel of 1, which indicates no overlapped patches. In Figure. 5.7, we observed similar scenario as high-resolution experiment, the mean DSC score is lower than hierarchical method. This effect is probably due to complete tiling act only on local patches, which lacks holistic context. The unsatisfied performance of the tiling patches presents two reasons. First, compare with hierarchical method, complete tiling contains unrelated patches without target. Second, the large variance of target present unbalanced intensity distribution. *Structural tiling vs random shift:* In this section, We evaluated translational data augmentation techniques. The right panel of Figure. 5.7 compares structural tiling, structural tiling plus random shift, and only random shift strategies. The structural tiling means adjacent patches cropped along axes such as complete tiling. With the increasing of average covered voxel, we shift the tiled patches to be overlapped by half,

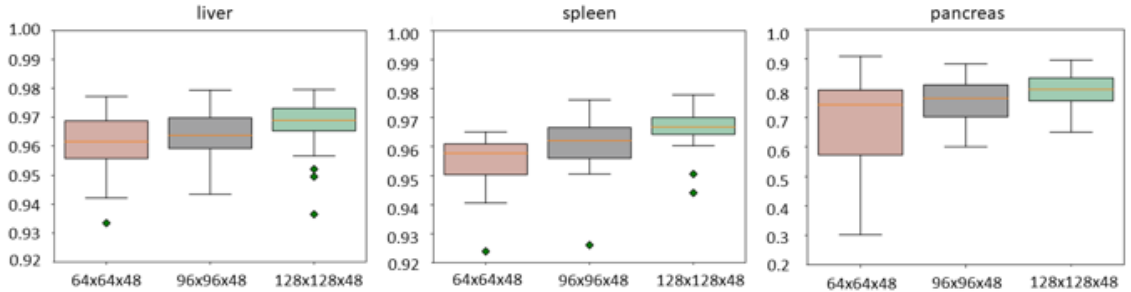


Figure 5.8: Boxplot on three different patch size along x-y axes. The ablation study conducted on three abdominal organs (spleen, liver and pancreas). Patch size range from small (64x64x48), medium (96x96x48) and large (128x128x48). The boxplots show that larger patch sizes perform better than smaller patch sizes.

Table 5.5: Average time cost per CT volume in the testing phase on different multi-organ segmentation models, where mean DSC is the average Dice score across 12 organs on BTCV testing data.

Methods	Mean DSC	Testing
Roth <i>et al.</i> (Roth et al., 2018b)	0.7920±0.065	304
Zhou <i>et al.</i> (Zhou et al., 2017)	0.8138±0.064	312
Zhu <i>et al.</i> (Zhu et al., 2018)	0.8328±0.063	294
Ours (N = 25)	0.8492±0.027	297
Ours (N = 50)	0.8564±0.060	308

1/3, 1/4, etc. The structured shifting was implemented in three dimensions (x, y and z axes) in order to balance spatial context for augmentation. The grey dash line in Figure. 5.7 indicates performance of structural tiling and random shift. From tiled patches, we implemented Gaussian random shift upon structural patches. This method involves moving the image randomly along the x, y, and z direction, which enables network to ignore absolute location of targets.

5.4.2.2 Effect of Average Number of Coverages per Voxel

We point out that patch-based approaches’ performance is partially influenced by number of overlapping region of interests. To pin-point the gain of increasing number of patches, we conducted large-scale of experiments on spleen by using coverage of 2 per voxel to 50 for each patch strategy. For structural tiling, the number of coverages per voxel is calculated by overlapping tiles. For random shifting, we count the mean coverages per voxel as the same with structural tiling. In Figure. 5.7. These competitors perform decently when number of coverages reach 10 and more. Curves represent the mean DSC of each experiment, while the shading area indicates variance. In-

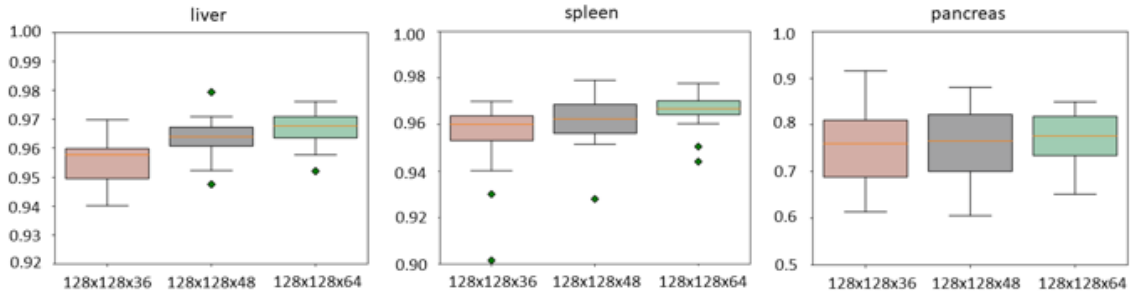


Figure 5.9: Boxplot on three different patch size along the z-axis. The experiments are conducted on spleen, liver and pancreas with patch size ranging from small (128x128x36), medium (128x128x48) and maximum (128x128x64). The boxplots also present that larger number of slices perform better than less in the volume.

Interestingly, the performance of random shift is higher than structural tiling when n is less than 15. For n larger than 15, the effect of random shift is less influential compared with pure structural tiling. Herein, we conclude that translational data augmentation is comparable to random shift effect when averaged patches per voxel reaches a large number.

5.4.2.3 Effect of Patch Size

Figure. 5.7 shows the results on dimension of 64/96/128. With the increasing of the scheduled dimension, all models perform better DSC, as indicated by larger spatial context. Big patches contain broader spatial context with consistent intensity distribution and trace of boundaries. Then, we conducted experiment on increasing of slice numbers. With the fixed x-y dimension of 128, we changed volume length from 36 to 64 shown in Figure. 5.7, it's no surprise that the performance follows similar result in x-y dimension. We conclude that 3D U-Net is capable to capture local features in patch with larger size, with the current computational resource, larger patch is better than small patches in segmentation metric.

5.4.2.4 Validation on External Datasets

In this evaluation, we investigated two clinical scenarios instead of research subjects. We adopt the model trained on the BTCV dataset and tested on external cohort with HEM1538 (splenomegaly) and ImageVU (pancreas). which are manually labeled by experts. Results are shown in Table. 5.2 and Figure. 5.10. *HEM1538*: The quantitative performance on HEM1538 is presented in Table. 5.2.

Table 5.6: Evaluation of different medical image segmentation methods on the BTCV testing dataset in multi-organ segmentation (12 organs). The evaluation is performed in terms of mean DSC and across organs.

2D methods		3D methods		Hybrid and 2.5D methods	
2D UNet (Ronneberger et al. 2015)	0.4935	3D UNet (Cicek et al. 2016)	0.5381	H-denseUNet (Li et al. 2018)	0.8172
ResNet (He et al. 2016)	0.5328	V-Net (Milletari et al. 2016)	0.5284	AH-Net. (Liu et al. 2018)	0.7947
Mask R-CNN (He et al. 2017)	0.7032	3D FCN (Chen et al. 2016)	0.5406	UMCT (Xia et al. 2018)	0.7984
DeepLab V3 (Chen et al. 2018)	0.8015	nnUNet (Isensee et al. 2018)	0.7934	OAN-RC (Wang et al. 2019)	0.7885
Ours (N=25)			0.8492		
Ours (N =50)			0.8564		

Table 5.7: Leaderboard of Multi-Atlas Abdomen Labeling Challenge (mean DSC).

Team	Mean Surface Distance	Average Dice	Hausdorff Distance
Try-1	1.3522	0.84056	20.3802
Try-2	1.4088	0.83626	20.0736
Path	2.9252	0.777832	32.6082
Ours	1.4237	0.85641	18.9862

We implemented the same comparison experiments with low-resolution, high-resolution and multi-scale hierarchy. The mean Dice similarity coefficient (DSC) is calculated for all testing scans in HEM1538. The multi-scale hierarchy method achieves best DSC among baseline approaches and was used as a reference method. Our proposed random patch network fusion models perform better than multi-scale hierarchy as presented in boxplot with significant improvement ($p < 0.01$, paired t-test). As HEM1538 is a pathology cohort with extra large spleens, we prove that RPFN could effectively preserves the stability on generalizing knowledge from normal spleen to splenomegaly. *ImageVU pancreas*: In this study, we introduced an outlier-guided cohort since clinically acquired scans contain hard cases among population. The quantitative result is presented in Table. 5.2, the mean volume DSC showed the detailed measurement for all methods, which showed that the proposed RPFN with 50 patches and majority vote achieves superior performance compared with the two-level multi-scale hierarchy with ($p < 0.01$, paired t-test with mean DSC).

5.4.3 Comparison of State-of-the-art Methods

5.4.3.1 Coarse-to-fine Methods

Our model is compared with other state-of-the-art coarse-to-fine networks. The results are in Table. 5.3. Roth et al. [66] used hierarchical 3D fully convolutional networks (FCN) with two stages. Zhou et al. [168] developed a fix-point model for small organ segmentation. Li et al. [178] combined 2D and 3D FCNs for hierarchically aggregating volumetric contexts. Zhu et al. [63] proposed

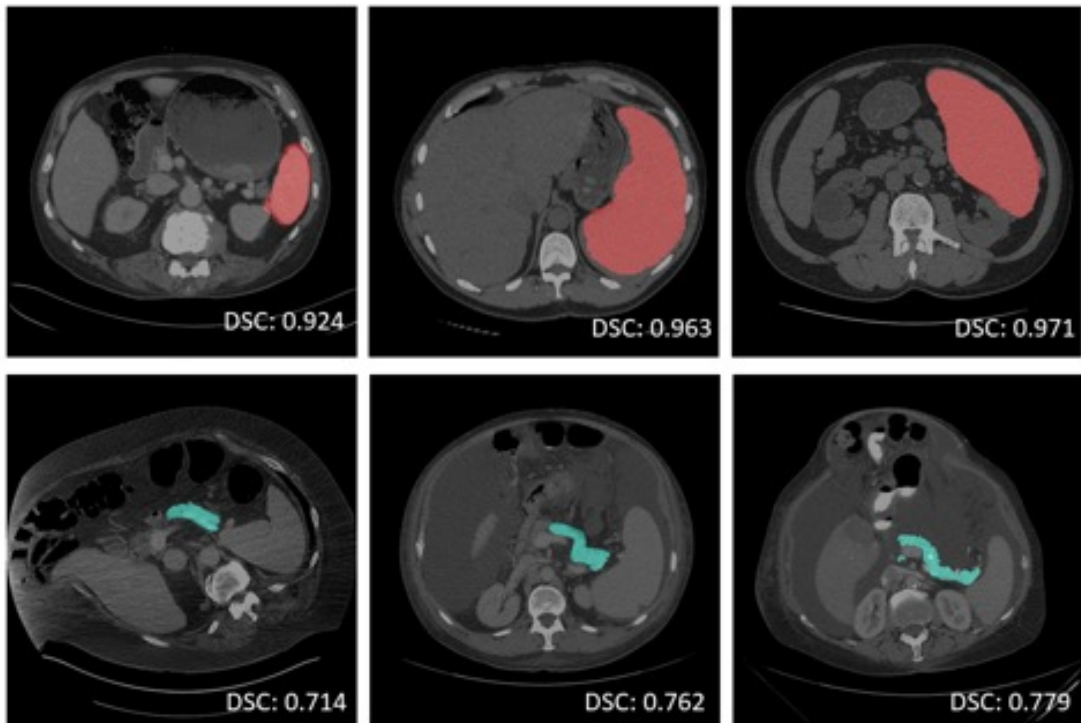


Figure 5.10: Qualitative result of three representative subjects. From low to high, we show the segmentation result evaluated by our method. The testing performance on external datasets (top: HEM1538-splenomegaly, bottom: ImageVU-pancreas outliers).

a novel 3D coarse-to-fine framework that achieved promising result on pancreas segmentation. For each method, we trained 12 models for 12 organs. Comparing with these state-of-the-art coarse-to-fine methods, our work achieves a consistent higher DSC. The average Dice of our method is 0.8564, compared to 0.7920 [66], 0.8138 [168], 0.8176 [178], 0.8328 [63], respectively.

5.4.3.2 Patch Selection Methods

We implemented different patch selection strategies used in abdominal organ segmentation. To fairly conduct the evaluation, we used 3D UNet as the segmentation model for all methods. All experiments are implemented using the same BTCV dataset on 12 organs. As shown in Table. 5.4, we compared five different strategies with our method in addition to the ablation study. The evaluation metric employed in these experiments includes the mean surface distance, Dice scores, and Hausdorff distance. In comparison of single-stage model, patches selected with overlap perform better than tiles without overlap. In comparison of two-stage model, we evaluated the fine-stage performance using fine-scaled, sliding window, and our random patch method. The average Dice of our method is 0.8564, compared to 0.8297, 0.7991, respectively.

As shown in Table. 5.4, our method outperforms other patch selection strategies in terms of three evaluation metrics. We observed that the fine-scaled and the sliding window method perform well when the first stage predictions are relatively good. But the fine-stage is sensitive to catastrophic failures predicted from coarse stage. The Zhu et al. [63] method outperforms Roth et al. [179] mainly because it involves the operation of expanding box ($n=12$). Our method achieved the improvement may be due to the smoothing effect introduced by random patches, it could save the catastrophic failures in the first stage. Figure. 5.5 shows visualization of sample results of our method compared to Roth et al. [179]. We could observe that the fixed patches may be vulnerable to the error field of view given by first-stage segmentation.

5.4.3.3 Comparison of Time Efficiencies with Different Methods

We discuss the average time cost of our proposed method against other coarse-to-fine methods. The number of patches used in the approach matters the overall testing time. Here, we choose $n = 50$ and $n = 25$ for discussing the concern of accuracy-time trade-off. In experiments of Roth et al. [179], Zhou et al. [168], and Zhu et al. [63], we trained 12 models for 12 abdominal organs. The time cost

evaluation is calculated after acquiring the final multi-organ segmentation output (including post-processing steps reported in each method). In implementing Zhu et al. [63]. We choose the overlap size $n = 6$ and 12 as noted in the study. Experimental results are shown in Table. 5.5. Zhu et al. [63] is the most efficient. Our method achieves comparable time efficiency on $N = 25$. When N is larger, the performance improves but the testing time also increases which is reasonable. We also observe that, in the testing phase of coarse-to-fine methods, the time of loading models composes the most part. Overall, coarse-to-fine methods take more than double seconds in the testing phase due to the loading of multiple models, and the automatic algorithms take much less time than radiologists, which presents the clinical significance of the work.

5.4.3.4 Comparison with Different Medical Image Segmentation Methods

We discussed different prevalent methods on medical image segmentation methods on the task of 12 abdominal organ segmentation. We used the same data split configuration during experiments. The results are shown in Table. 5.6. In comparison of 2D methods, the basic model 2D UNet [26] and ResNet [24] suffer from worse DSC of small organs such as adrenal glands and gallbladder. Mask R-CNN [25] and DeepLab V3 [180] achieves higher performance because the localization effect in the framework. In experimental results of 3D methods. We observe low performance due to the severely down-sampled volume of CT images in 3D UNet [115], V-Net [156], and 3D FCN [181], While nnUNet [58] benefits from the cascaded framework that incorporates many ensembled predictions in the outputs. In comparison of 2D/3D hybrid networks, former methods achieve comparable results, these studies utilized both 2D and 3D context in a single network. Compare with above current state-of-the-art medical image segmentation methods, coarse-to-fine frameworks achieve consistent higher DSC in the task, probably due to the effective combination of low-resolution context and high-resolution contexts.

5.4.3.5 Comparison with Multi-Atlas Abdomen Labeling Challenge leaderboard

The result of top teams on the leaderboard are listed in Table. 5.7. Note that “try” team leads the top several rankings. It is also noticeable that some latest work achieved high performance such as Zhou et al. [44] that achieves mean DSC of 0.850 are not reported on the leaderboard. Compare with the leaderboard, our method outperforms other state-of-the-art methods, and achieve the highest mean

DSC and the best Hausdorff distance performance in the standard competition.

5.5 Conclusion and Discussion

In the work, we revisit the challenging whole volume based 3D abdominal segmentation. Due to limitations in low-resolution, high-resolution and hierarchy approaches under restricted GPU memory, we explored the usage of randomly selected patches to the hierarchical method. First, we provided a 3D coarse multi-organ segmentation using 3D U-Net. Then, we implemented the random sampling to crop the context around target for removing fixed pattern in patches. Next, we trained a high resolution fine-tuning network to compensate the shape and boundary structure for patches. Last, we employed majority vote mechanism to fuse the full segmentation mask for CT scan. Moreover, we conducted exhaustive experiments on comparison of different strategies of patch-based methods, we demonstrated that translational data augmentation and random sampling both provided boosted performance in comparison to simply adding more patches in 3D CNN. Additionally, we did large scale of experiments on effect of average covered patches per voxel, which we conclude that more patches generously perform better than a smaller number of patches. However, majority vote works differently on variant structures, too many patches only increase the computational time in some anatomies. Besides, we deployed the trained models on two unseen datasets, we show that the method can be generalized to outlier cases and pathological cohort.

In this study, the proposed random patch network fusion enables the training to address the memory issue for high dimensional 3D abdominal segmentation. In this study, 50 random patches are used for segmenting variant isotropic resolution at $(0.8 \times 0.8 \times 2)$ for CT scans. It could be possible that GPU memory would be large enough for housing entire abdomen CT in the future. However, the local-global feature trade-off games would still exist. Herein, the random patch network fusion technique could be a good choice for such scenarios.

The major limitation of the proposed method is that the computational time would be linearly accumulated with the increasing number of random patches. Besides, the majority vote algorithm is not time efficient when applied to voxel-wise voting. Another disadvantage of majority vote is that when voxel is rarely covered by voters, the voxel is vulnerable to be miss-labeled. Therefore, it's appealing to investigate better statistical fusion algorithm with more efficient time and space complexity, while perverse stability for removing outlier labels.

Another limitation in this work is that the hierarchical labeling framework still failed to mimic doctor's process for identifying structures. In the future, the hierarchical labeling could be investigated in clinical inspired approaches. Instead of simply transferring low-resolution feature to high resolution model, we could also pass anatomies' features to next hierarchy. For example, radiologist would first find portal and splenic vein before identifying pancreas. If we could transfer the correlated feature as a prior in different levels' hierarchy, the performance would be potentially improved. In addition, previous works indicated a high-dimensional data often suffer redundancy (e.g., not every voxel in 3D volume is useful). Mining the boundary of organ versus backgrounds or other tissues could leads to a more efficient model. We posit that it is worth additional study of the patch selection strategy around boundaries. Future work could study the efficacy of boundary patches and inner voxel patches.

In summary, the proposed random patch network fusion achieved consistent superior segmentation performance compare with other labeling frameworks, since it led to a better balance between performance and computational cost compared to other patch-based and multi-stage approaches. The balanced configurations are fulfilled by introducing 1) two-stage hierarchical levels, 2) randomly localized patches, 3) network label fusion. Our method presented positive result of 3D abdominal segmentation in variety of structures and datasets. We hope random patch network fusion will be useful with other context tasks that involve hierarchical labeling design.

CHAPTER 6

Pancreas CT Segmentation by Predictive Phenotyping

6.1 Introduction

Patient care data, such as CT scans and electronic health records (EHR) with the pancreatic disease, are heterogeneous in nature. Disease progression and treatment delivery are associated with different care trajectories, which in turn lead to varying pancreas patterns. In anticipation of diabetes, patients are observed with atrophic pancreas tissues [182], with progression noted in the patients' medical history in terms of International Classification of Diseases (ICD) codes. Current pancreas segmentation methods [43, 183, 184, 185] are typically driven by imaging data, while phenotype covariates [186, 187, 188, 189, 190] that indicate underlying patient conditions are not well considered. We observe that different disease types present heterogeneous textures (Fig. 6.1), and thereby hypothesize that identifying different pancreas patterns can extract the discriminative contexts which can well benefit pancreas segmentation.

Data-driven phenotype clustering has been recently used to group patients sharing close outcomes [191, 192, 193, 194]. Combining imaging biomarkers, Virostko *et al.* [195] assessed the pancreas size with type I diabetes patients. Tang *et al.* [121] showed the feasibility of onset type II diabetes prediction using CT scans. However, to date, how to fully exploit the EHR data for guiding medical image segmentation has been rarely studied. A naïve approach is to simply concatenate both image and EHR data as a two-channel input, and then train a standard convolutional neural network for deriving the outcome. However, this fusion strategy is not directly applicable for our task since 1) a patient can have hundreds of phenotype categories; 2) the fusion strategy cannot account for patients' observed outcomes (*e.g.*, onset of comorbidities, chronic progression of metabolic syndrome); and 3) it requires EHR data as input during inference, which does not commonly exist for many real-world pancreas segmentation datasets (*e.g.*, BTCV MICCAI Challenge [72], TCIA pancreas CT [196]).

To address above challenges, we propose the first pancreas segmentation framework to model both the pancreas imaging features and clinical features via predictive phenotyping. The rationale

is that the larger scale of EHR data with (*e.g.*, ICD-10 code) which indicates phenotype subgroups can be potentially correlated to the different appearance of the pancreas. Specifically, the proposed approach consists of an encoder, a segmentation decoder, and a predictor with sets of phenotypes candidates’ centroids. Our method is designed to meet the following requirements: 1) The subject image should be partitioned into several subgroups sharing similar future outcomes; 2) The assigned discrete representation should retain the patient phenotype context, and 3) The phenotype representation is used as prior knowledge for predicting. Particularly, in our framework, the encoder maps an image into a latent representation; the predictor assigns one or several phenotype categories by taking the latent variable as input; the segmentation decoder estimates the pixel-wise labels conditioned on the assigned centroid. To homogenize future outcomes in each subgroup, we introduce a phenotyping objective given CT images by regularizing Kullback-Leibler (KL) divergence between the learned latent representation and the embedding centroid. Finally, the segmentation model estimates the pancreas segmentation mask given encoding of the image and the risk embedding.

Our contribution is four-folds: we successfully (1) learn a phenotype embedding between CT and EHR; (2) formulate a pancreas segmentation framework that benefits from predicting phenotype subgroups; (3) demonstrate improved pancreas segmentation performance on healthy and disease patient cohorts; (4) design the embedding approach without requiring EHR at the testing phase. The significance of the study is that we use an experimental CT imaging-phenotyping approach for investing clinical underpinnings of pancreas segmentation. The phenotype embedded model enriches segmentation contexts improving the characterization of heterogeneous disease and allows for deeper consideration of patient phenotype in image-based learning.

6.2 Method

6.2.1 Problem Formulation

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be variables for input images and an output segmentation label. $C \in \mathcal{C}$ is the patient phenotype onset (*i.e.*, one or a combination of future outcomes) where \mathcal{X} , \mathcal{Y} and \mathcal{C} are the image feature, label, and phenotype onset space, respectively. Specifically, CT image X selection is censored from timestamp: date of diagnostic code is later than date of scan at least 1 year. The input of C is the sequence of covariate admissions, the feature of one admission is a multi-hot vector containing the comorbidities or demographics: $C'_{EHR} = [M', M_1, M_2, \dots, M_\ell]$. M' is

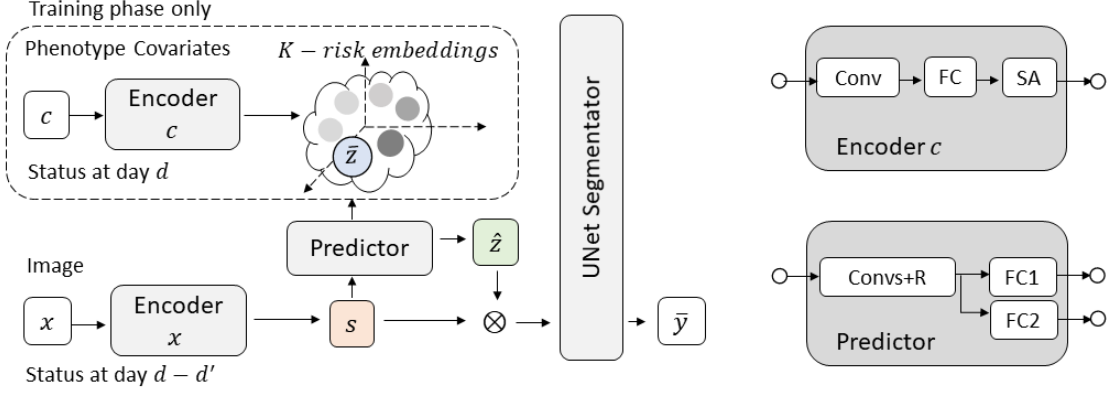


Figure 6.1: Phenotype embedded segmentation architecture in the training phase. The left diagram shows the embedding network combining image features, predictive phenotyping, pre-existing risk conditions lying in the latent space to be fed into the segmentation model. The predictor is trained for predicting phenotype-dependent feature maps and selecting “similar” cluster assignment, where the phenotype information is not required as input in the testing phase. Right top: encoder for processing phenotype covariates, right bottom: the predictor follows self-training scheme from image feature. Here, SA denotes soft assignment for risk embeddings, R for ReLU, FC for fully connected layers.

the set of demographic values, M_C is the set of phenotype admissions constructed by binary vectors of aggregate ICD-10 codes. In the training phase, we are given the dataset $\mathcal{D} = \{x^n, y^n, c^n\}_{n=1}^N$ consists of observations (x, y, c) for N subjects. In the testing phase, we assume the dataset only comprising image volume $\{x^n\}_{n=1}^N$. The goal is first to identify a set of K predictive phenotypes $\mathcal{L} = \{z_1, z_2, \dots, z_K\}$ lying in the latent space. Each phenotype cluster is supposed comprising of homogeneous patients that can be represented by the cluster centroid. This predictive phenotyping updates the encoder to suggest the context to which cluster a patient belongs. Second, we design the segmentation model to estimate the pixel-wise label given the encoding variable and the predictive phenotyping distribution. Let \bar{z} be the random variable that lying in the phenotype onset latent space and s be the image feature. The predictive phenotyping can be fulfilled by optimizing the Kullback-Leibler (KL) divergence between distributions conditioned on the image: $p(\hat{z}|s)$ and onset phenotypes $p(\bar{z}|c)$, respectively. Combining the aim of segmentation, we establish our goal as following objective:

$$\mathbb{E}_{Y \sim (x, y, c)} [-\log P(y|s, \hat{z}_k)] + KL(\hat{z}_k|s \parallel \bar{z}_k|c). \quad (6.1)$$

6.2.2 Loss Functions

Loss functions are designed to meet the objective in Eq. 6.1 and are proposed to iteratively refine the predicted phenotyping from image features. Specifically, our model is trained by matching image distribution to the target distribution defined by future outcomes. To this end, we define the objective as KL divergence between an expectation and the cluster assignment:

$$\mathcal{L}_1(\bar{z}, \hat{z}) = \mathbb{E}_{z \sim P(x,c)} \left[- \sum_{k=1}^K \bar{z}_k \log \hat{z}_k \right], \quad (6.2)$$

where \bar{z}_k and \hat{z}_k indicate the k -component of \bar{z} , \hat{z} , respectively. Note that the KL divergence loss reaches its minimum when two latent distributions are equivalent. Additionally, the segmentation loss penalizes the predicted mask \bar{y} and the ground truth label y by DSC-loss:

$$\mathcal{L}_2(\bar{y}, y) = \mathbb{E}_{Y \sim P(s,\hat{z})} \left[1 - \frac{2 \times \sum_i y_i \bar{y}_i}{\sum_i y_i + \sum_i \bar{y}_i} \right], \quad (6.3)$$

where the form follows [197] to prevent a model from background bias.

6.2.3 Phenotype Embedding

To encourage homogeneous future outcomes in each phenotyping cluster, we employ embedded mapping [198] as our initialization method. Given an initial estimate of the non-linear mapping c' and cluster centroid μ . We adopt the self-supervision [199] training strategy that iteratively 1) optimizes soft-assignment between embedded points and clustering centroids; 2) updates deep mapping and centroids. The soft-assignment block in Fig. 6.1 follows [200] using Student's t -distribution as a kernel to estimate between data points and cluster centroid:

$$q_{ik} = \frac{\exp(1 + \|c'_i - \mu_k\|^2 / \alpha)}{\sum_j \exp(1 + \|c'_i - \mu_j\|^2 / \alpha)}, \quad (6.4)$$

where α denotes the degrees of freedom of Student's t -distribution ($\alpha = 1$ for all experiments), \exp is the exponential operation with power $-(\alpha + 1)/2$ and q_{ik} can be interpreted as the probability of assigning sample i to cluster k . More comparisons of clustering benchmarks can be found in [193, 194]. After initialization, the embedding learning is iteratively updated during segmentation training.

6.3 Experiments

6.3.1 Dataset

The abnormal pancreas segmentation dataset. We have curated an abnormal pancreas dataset that contains 2000 adult patients (aged 18-50 years) with 14927 recorded visits and de-identified longitudinal CT scans under IRB approval. Each patient is associated with 101 covariates under radiologists’ query, including information on demographic and abdomen-related comorbidities that can potentially impact pancreas tissues. CT images are acquired at least one year (range from 1.0 to 2.1 years) earlier than diagnosis codes for each patient, to meet the requirements of the prognostic task with predictive phenotyping. For the segmentation task, 300 patients’ CT images are annotated and used for experiments. Each CT scan is $512 \times 512 \times \text{Slices}$, where the number of slices ranges from 72 to 121 under the body part regression process of [125] to acquire relatively same abdomen region of interest (ROI). The slice thickness ranges from 1mm to 2.5mm.

BTCV MICCAI Challenge 2015. We used the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge [72] as one of the external testing sets. The challenge dataset contains 50 abdominal CT scans. For evaluating the testing phase of the proposed method, the dataset does not include patient phenotype information. Each CT scan is manually labeled with 13 structures including pancreas with a spatial resolution of $([0.54 \sim 0.54] \times [0.98 \sim 0.98] \times [2.5 \sim 5.0] \text{ mm}^3)$.

TCIA pancreas CT. We use the 82 abdominal contrast enhanced CT scans from National Institutes of Health Clinical Center as the second external testing set. The publicly available study cohort contains 17 kidney donor subjects, and 65 patients were selected with no pancreatic cancer lesions and pathology. Each CT scan is in a resolution of 512×512 and slice thickness of $[1.5 \sim 2.5] \text{ mm}$.

6.3.2 Implementation Details

Follow prevailing pancreas segmentation baselines [43, 184, 185], we adopt the coarse-to-fine strategy for 3D pancreas segmentation. The coarse stage takes a highly down-sampled CT volume at an input dimension of $164 \times 164 \times 64$. For the fine stage, we cropped $64 \times 64 \times 64$ sub-volumes constrained to be in the pancreas region of interest (ROI). For experiments, 10% and 20% of subjects are randomly selected as validation and testing sets with the in-house dataset. Note that, the two external datasets are only used for testing, no subjects are used for the training procedure. We used 1)

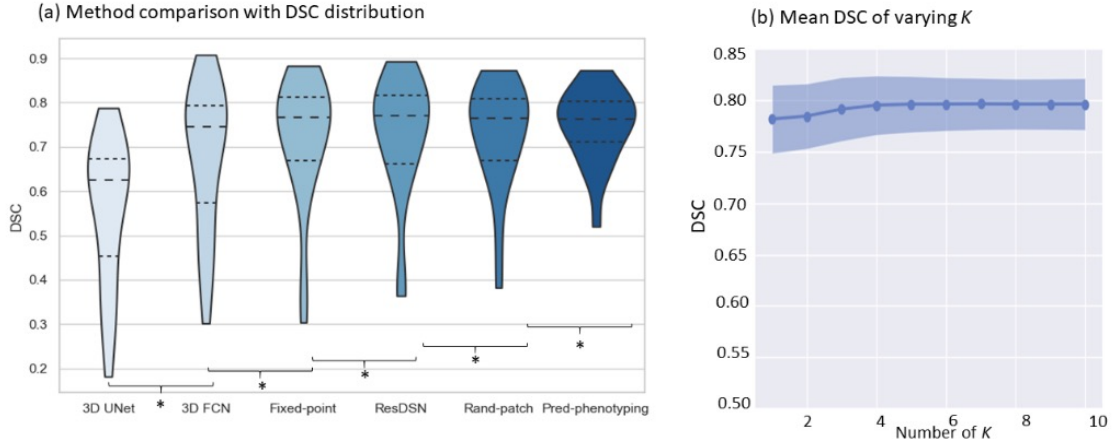


Figure 6.2: Testing performance on ImageVU dataset. Left: Distribution (median and quartiles) of DSC, the predictive phenotyping shows smaller variance and reduces the number of outliers (DSC<0.4). Right: The DSC (mean) comparison with varying K . The performance shows higher improvement as K increase from 1 to 4, then becomes marginal after $K = 4$. * denotes statistically significant under Wilcoxon signed-rank test ($p < 0.05$).

CT window range of $[-175, 275]$ HU; 2) scaled intensities of $[0.0, 1.0]$; 3) training with Nvidia 2080 11GB GPU with Pytorch implementation; 4) Adam optimizer with momentum 0.9. The Learning rate is initialized to 0.001 followed by a factor of 10 every 50 epochs decay.

Metrics. Segmentation performance is evaluated between ground truth and prediction by Dice-Sorensen coefficient (DSC), Averaged Surface Distance (ASD), and symmetric Hausdorff Distance (HD).

6.3.3 Comparison with State-Of-The-Arts

We compare the proposed method with various state-of-the-art methods: 1) 3D-UNet[115]; 2) hierarchical 3D FCN [201] (denoted as “3D FCN”); 3) the fixed-point model [184] (denoted as “C2F Fixed-point”); 4) 3D ResDSN [185] (denoted as “C2F ResDSN”); and 5) the random patches model [43] (denoted as “C2F Random-patches”). Here “C2F” denotes the coarse-to-fine training strategies [43, 184, 185].

6.3.4 Results

We compare our method against state-of-the-art approaches with respect to the cluster number at 4 (Table6.1). Our method significantly improves performance in terms of DSC, ASD and HD, with

Table 6.1: Performance comparison on the abnormal pancreas segmentation dataset. C2F denotes coarse-to-fine training. * denotes statistically significant against above method with Wilcoxon signed-rank test.

Methods	DSC	ASD	HD
3D-UNet (Cicek <i>et al.</i>)	0.697	5.592	27.154
3D FCN (Holger <i>et al.</i>)	0.724*	4.042*	25.195*
C2F Fixed-point (Zhou <i>et al.</i>)	0.746*	2.981*	22.516*
C2F ResDSN (Zhu <i>et al.</i>)	0.767*	2.105	22.017
C2F Random-patches (Tang <i>et al.</i>)	0.775*	1.976*	20.591*
Predictive phenotyping (Ours, K=4)	0.791*	1.697*	19.482*

Table 6.2: External testing performance comparison on BTCV MICCAI Challenge 2015 and TCIA pancreas (mean DSC) with our model trained on the internal data. Note that no subject from these two datasets are used for training. C2F denotes coarse-to-fine training strategies. * for statistically significant against above method with Wilcoxon signed-rank test.

Methods	BTCV	TCIA
3D-UNet (Cicek <i>et al.</i>)	0.685	0.770
3D FCN (Holger <i>et al.</i>)	0.709*	0.776*
C2F Fixed-point (Zhou <i>et al.</i>)	0.726*	0.797*
C2F ResDSN (Zhu <i>et al.</i>)	0.730*	0.804
C2F Random-patches (Tang <i>et al.</i>)	0.742*	0.813*
Predictive phenotyping (Ours, K=4)	0.757*	0.822*

$p < 0.05$ under Wilcoxon signed-rank test. Importantly, the Hausdorff distance (HD) improvement shows that EHR information provided useful context to reduce outliers. In Table 6.2, we further investigate the comparison experiment results with external testing sets. The two public challenge data do not include patient EHR, *i.e.*, demographics, ICD codes. Our method implicitly predicts the future outcomes from the image feature and fused to the segmentation task. The method achieves a mean DSC of 0.757 on BTCV data, and 0.827 on TCIA pancreas CT. Predictive phenotyping improves several outlier cases, showing less variance (Fig 6.2). Qualitative inspection confirms the numerical results (Fig 6.3). First, we inspect the data of a patient with potential lung infections and relatively normal pancreas tissue. In the second case, the patient has type I diabetes, observing a degraded pancreas tissue. Importantly, the improvement with respect to the degraded pancreas is larger than the healthy pancreas, showing the predictive phenotyping can be informative for identifying variant patterns.

Ablative Study. *Efficacy of the predictive phenotyping and network architecture* In Table 1, we compared the backbone model (row 5) and predictive phenotyping (row 6). The EHR improved performance on two datasets by 1.5%, significant Wilcoxon signed-rank test, $p < 0.001$. For the dia-

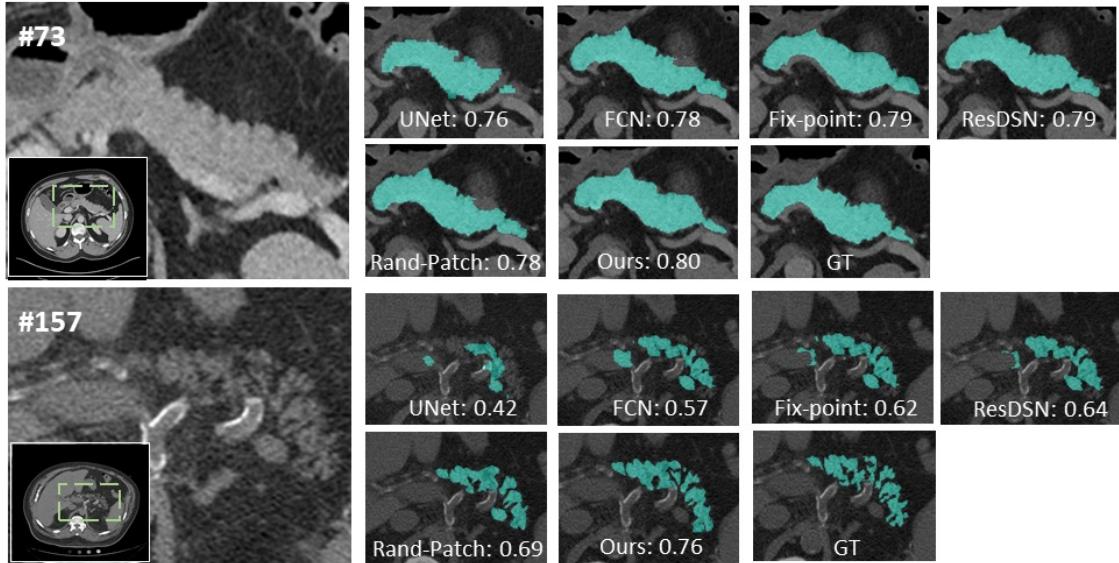


Figure 6.3: Two representative cases. The top subject has potential lung infections and relative normal pancreas tissue. The bottom case has type 1 diabetes with degraded pancreas tissue. The accuracy gain of the diabetes case is larger than the normal case, showing the method’s ability for identifying degraded pancreas tissue.

betic patients, performance improvement gains were larger. Predictive phenotyping with the EHR outperforms naïve approach with feature concatenation by a large margin, from 74.5% to 77.9%. We compared with pancreas segmentation state-of-the-art methods. Predictive phenotyping significantly improved performance in terms of DSC, ASD and HD, with $p < 0.05$, Wilcoxon signed-rank test. Importantly, HD improvement shows that EHR information provided useful context to reduce outliers. We further investigate the comparison experiment results with external testing sets. For external validation, the two public challenge data do not include patient EHR, i.e., demographics, ICD codes. Our method implicitly predicts the future outcomes from the image feature and fused to the segmentation task. The method achieves a mean DSC of 0.757 on BTCV data, and 0.827 on TCIA pancreas CT.

Importance of hyper-parameter K We further evaluate the performance by varying the number of clusters K from 1 to 10 on the in-house dataset. Fig. 6.2 shows improved DSC as K increased, the DSC improves from 0.7814 to 0.7956 as K from 1 to 4. The performance is observed no significant improvement after $k = 4$ ($p < 0.1$, Wilcoxon signed-rank test).

6.4 Discussion and Conclusion

Comparing Table 1 and Table 2, the proposed method shows higher improvement over baseline methods if the cohort has more severe cases of abdominal diseases. Specifically, the performance improvement on the abnormal pancreas segmentation dataset in terms of the average Dice is 1.6%, which is larger than that of the BTCV dataset (1.5%), and the TCIA dataset (1.1%), respectively. We have also demonstrated two qualitative examples. To show that our method can lead to more performance gain for the atrophic pancreas than the normal pancreas. The larger improvement on diseased cohort can be a potential advantage of the phenotype embedding. In addition to the major segmentation objective, the case-specific feature projected to the phenotype embedding space can be observed. The comorbidities developed in the next two years with 4 identified clusters, and listed ICD-10 codes are with most frequencies in each grouped phenotype component. The first component shares the most cases with relative normal pancreas, while the second, third and fourth indicate varying phenotype outcomes of the atrophic pancreas, metabolic syndrome, and pancreas with inflammatory fats, respectively. The number of phenotype components K is one of the most important parameters in the study: increasing k can potentially impact the predictive embedding with higher diversity representing data distribution. However, the interpretability will decrease as it shares fewer similar data points. In the future, the interpretability of the predicted phenotyping can be further evaluated with more clinically meaningful investigations.

In this work, we introduce pancreas segmentation by predictive phenotyping, a patient-oriented approach for understanding between EHR and CT data. The experimental imaging-phenotyping approach is used for investigating the phenotype underpinnings of the pancreas. We demonstrate a predictive task to encourage image embedding to the phenotyping cluster with similar patient outcomes. The EHR data is designed as input at the training phase, and only images are required for inferring phenotyping context and segmentation at the test phase. Throughout experiments on the in-house dataset and two public challenge datasets, we show that the method highlights a significant role over state-of-the-art segmentation. The integrated imaging-phenotyping method could encourage solutions that better respect anatomical variability, especially associated with disease progression or comorbidities. When EHR data is available, the method can be applied for boosting performance.

CHAPTER 7

Spatial Long-Range Dependencies: Transformers for 3D Medical Image Segmentation

7.1 Introduction

Vision transformers have recently gained traction for computer vision tasks. Dosovitskiy et al. [88] demonstrated state-of-the-art performance on image classification datasets by large-scale pre-training and fine-tuning of a pure transformer. In object detection, end-to-end transformer-based models have shown prominence on several benchmarks [202, 203]. Recently, hierarchical vision transformers with varying resolutions and spatial embeddings [204, 205, 206, 207] have been proposed. These methodologies gradually decrease the resolution of features in the transformer layers and utilize sub-sampled attention modules. Recently, multiple methods were proposed that explore the possibility of using transformer-based models for the task of 2D image segmentation [90, 208, 209, 210]. Zheng et al. [90] introduced the SETR model in which a pre-trained transformer. We introduce a novel architecture, dubbed as UNet TRansformers (UNETR) shown in Figure. 7.1, that utilizes a transformer as the encoder to learn sequence representations of the input volume and effectively capture the global multi-scale information, while also following the successful “U-shaped” network design for the encoder and decoder.

7.2 Data

Medical Segmentation Decathlon (MSD) dataset [211] comprises of 10 segmentation tasks from different organs and image modalities. These tasks are designed to feature difficulties across medical images, such as small training sets, unbalanced classes, multi-modality data and small objects. Therefore, the MSD challenge can serve as a comprehensive benchmark to evaluate the generalizability of medical image segmentation methods. The pre-processing pipeline for this dataset is outlined in supplementary materials. BTCV : The Beyond the Cranial Vault (BTCV) abdomen challenge dataset [72] contains 30 subjects with abdominal CT scans where 13 organs are annotated by interpreters under supervision of radiologists at Vanderbilt University Medical.

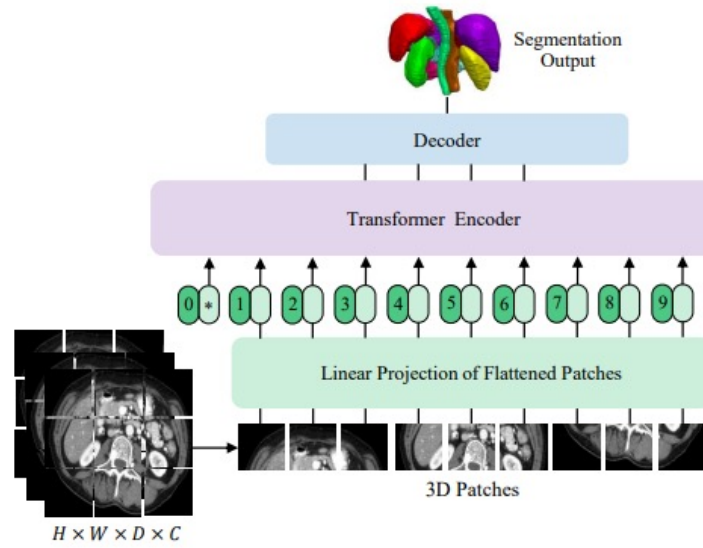


Figure 7.1: Overview of UNETR. Our proposed model consists of a transformer encoder that directly utilizes 3D patches and is connected to a CNN-based decoder via skip connection.

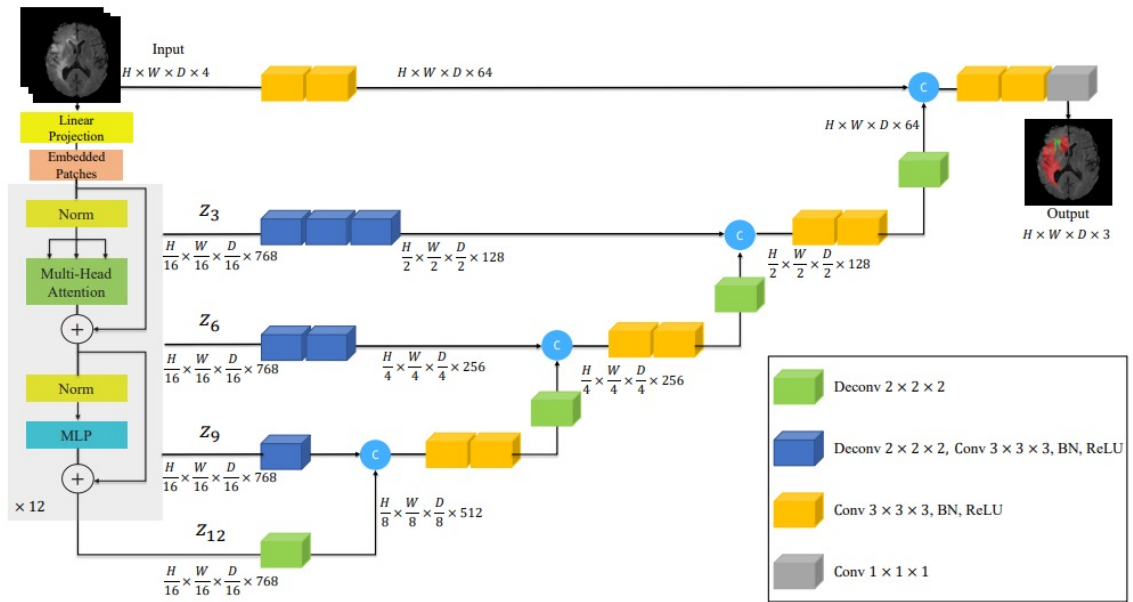


Figure 7.2: Overview of UNETR architecture. A 3D input volume (e.g. $C = 4$ channels for MRI images), is divided into a sequence of uniform non-overlapping patches and projected into an embedding space using a linear layer. The sequence is added with a position embedding and used as an input to a transformer model. The encoded representations of different layers in the transformer are extracted and merged with a decoder via skip connections to predict the final segmentation. Output sizes are given for patch resolution $P = 16$ and embedding size $K = 768$.

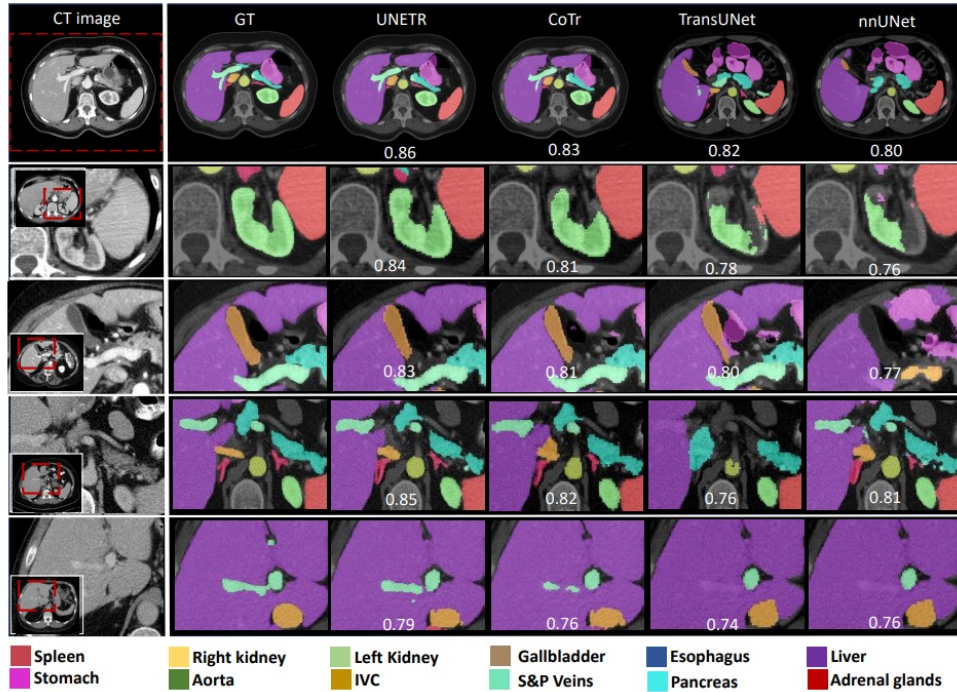


Figure 7.3: Qualitative comparison of different baselines in BTCV cross-validation. The first row shows a complete representative CT slice. We exhibit four zoomed-in subjects (row 2 to 5), where our method shows visual improvement on segmentation of kidney and spleen (row 2), pancreas and adrenal gland (row 3), gallbladder (row 4) and portal vein (row 5). The subject-wise average Dice score is shown on each sample.

7.3 Experiment

We propose to leverage the power of transformers for volumetric medical image segmentation and introduce a novel architecture dubbed as UNet Transformers (UNETR). We reformulate the task of 3D segmentation as a 1D sequence-to-sequence prediction problem and use a transformer as the encoder to learn contextual information from the embedded input patches. The extracted representations from the transformer encoder are merged with the CNN-based decoder via skip connections at multiple resolutions to predict the segmentation outputs. Instead of using transformers in the decoder, our proposed entire framework (shown in Figure. 7.2) uses a CNN-based decoder. This is due to the fact that transformers are unable to properly capture localized information, despite their great capability of learning global information.

7.4 Results

UNETR outperforms the state-of-the-art methods for both Standard and Free Competitions on the BTCV leaderboard. As shown in Table VII-1, in the Free Competition, UNETR achieves an overall average Dice score of 0.899 and outperforms the second, third and fourth top-ranked methodologies by 1.238%, 1.696% and 5.269% respectively. In the Standard Competition, we compared the performance of UNETR against CNN and transformer-based baselines. UNETR achieves a new state-of-the-art performance with an average Dice score of 85.3% on all organs. Specifically, on large organs, such as spleen, liver and stomach, our method outperforms the second best baselines by 1.043%, 0.830% and 2.125% respectively, in terms of Dice score. Furthermore, in segmentation of small organs, our method significantly outperforms the second best baselines by 6.382% and 6.772% on gallbladder and adrenal glands respectively, in terms of Dice score.

7.5 Discussion

Our experiments in all datasets demonstrate superior performance of UNETR over both CNN and transformer-based segmentation models. Specifically, UNETR achieves better segmentation accuracy by capturing both global and local dependencies. In qualitative comparisons, this is illustrated in various cases in which UNETR effectively captures long-range dependencies (e.g. accurate segmentation of the pancreas tail in Figure. 7.3). Moreover, the segmentation performance of UNETR on the BTCV leaderboard demonstrates new state-of-the-art benchmarks and validates its effectiveness. Specifically for small anatomies, UNETR outperforms both CNN and transformer-based models. Although 3D models already demonstrate high segmentation accuracy for small organs such as gallbladder, adrenal glands, UNETR can still outperform the best competing model by a significant margin. This is also observed in Figure. 7.3, in which UNETR has a significantly better segmentation accuracy for left and right adrenal glands, and UNETR is the only model to correctly detect branches of the adrenal glands. For more challenging tissues, such as gallbladder in row 4 and portal vein in row 5, which have low contrast with the surrounding liver tissue, UNETR is still capable of segmenting clear connected boundaries. This section introduces a transformer-based architecture for semantic segmentation of volumetric medical images by defining the task as sequence-to-sequence prediction problem. We proposed to use a transformer encoder to increase the model’s capability for learning long-range dependencies and effectively capturing global contextual

representation at multiple scales.

CHAPTER 8

Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis

8.1 Introduction

Vision Transformers (ViT)s [88] have started a revolutionary trend in computer vision [212, 213] and medical image analysis [94, 214]. Transformers demonstrate exceptional capability in learning pre-text tasks, are effective in learning of global and local information across layers, and provide scalability for large-scale training [215, 216]. As opposed to convolutional neural networks (CNNs) with limited receptive fields, ViTs encode visual representations from a sequence of patches and leverage self-attention blocks for modeling long-range global information [216]. Recently, Shifted windows (Swin) Transformers [204] proposed a hierarchical ViT that allows for local computing of self-attention with non-overlapping windows. This architecture achieves linear complexity as opposed to quadratic complexity of self-attention layers in ViT, hence making it more efficient. In addition, due to the hierarchical nature of Swin Transformers, they are well-suited for tasks requiring multi-scale modeling.

In comparison to CNN-based counterparts, transformer-based models learn stronger features representations during pre-training, and as a result perform favorably on fine-tuning downstream tasks [216]. Several recent efforts on ViTs [217, 218] have achieved new state-of-the-art results by self-supervised pre-training on large-scale datasets such as ImageNet [28].

In addition, medical image analysis has not benefited from these advances in general computer vision due to: (1) large domain gap between natural images and medical imaging modalities, like computed tomography (CT) and magnetic resonance imaging (MRI); (2) lack of cross-plane contextual information when applied to volumetric (3D) images (such as CT or MRI). The latter is a limitation of 2D transformer models for various medical imaging tasks such as segmentation. Prior studies have demonstrated the effectiveness of supervised pre-training in medical imaging for different applications [219, 220]. But creating expert-annotated 3D medical datasets at scale is a non-trivial and time-consuming effort.

To tackle these limitations, we propose a novel self-supervised learning framework for 3D med-

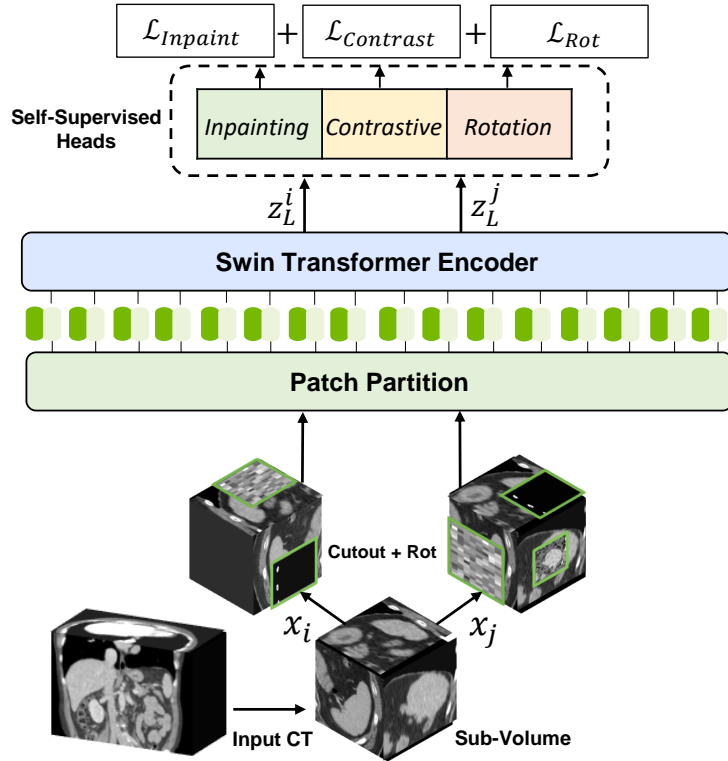


Figure 8.1: Overview of our proposed pre-training framework. Input CT images are randomly cropped into sub-volumes and augmented with random inner cutout and rotation, then fed to the Swin UNETR encoder as input. We use masked volume inpainting, contrastive learning and rotation prediction as proxy tasks for learning contextual representations of input images.

ical image analysis. First, we propose a new architecture dubbed Swin UNETR Transformers (Swin UNETR) with a Swin Transformer encoder that directly utilizes 3D input patches. Subsequently, the transformer encoder is pre-trained with tailored, self-supervised tasks by leveraging various proxy tasks such as image inpainting, 3D rotation prediction, and contrastive learning (See Fig. 8.1 for an overview). Specifically, the human body presents naturally consistent contextual information in radiographic images such as CT due to its depicted anatomical structure [124, 125]. Hence, proxy tasks are utilized for learning the underlying patterns of the human anatomy. For this purpose, we extracted numerous patch queries from different body compositions such as head, neck, lung, abdomen, and pelvis to learn robust feature representations from various anatomical contexts, organs, tissues, and shapes.

Our framework utilizes contrastive learning [221], masked volume inpainting [74], and 3D rotation prediction [222] as pre-training proxy tasks. The contrastive learning is used to differentiate

various ROIs of different body compositions, whereas the inpainting allows for learning the texture, structure and correspondence of masked regions to their surrounding context. The rotation task serves as a mechanism to learn the structural content of images and generates various sub-volumes that can be used for contrastive learning. We utilize these proxy tasks to pre-train our proposed framework on a collection of 5,050 CT images that are acquired from various publicly available datasets.

Furthermore, to validate the effectiveness of pre-training, we use 3D medical image segmentation as a downstream application and reformulate it as a 1D sequence-to-sequence prediction task. For this purpose, we leverage the Swin UNETR encoder with hierarchical feature encoding and shifted windows to extract feature representations at four different resolutions. The extracted representations are then connected to a CNN-based decoder. A segmentation head is attached at the end of the decoder for computing the final segmentation output. We fine-tune Swin UNETR with pre-trained weights on two publicly available benchmarks of Medical Segmentation Decathlon (MSD) and the Beyond the Cranial Vault (BTCV). Our model is currently the state-of-the-art on their respective public test leaderboards.

Our main contributions in this work are summarized as follows:

- We introduce a novel self-supervised learning framework with tailored proxy tasks for pre-training on CT image datasets. To this end, we propose a novel 3D transformer-based architecture, dubbed as Swin UNETR, consisting of an encoder that extracts feature representations at multiple resolutions and is utilized for pre-training.
- We demonstrate successful pre-training on a cohort of 5,050 publicly available CT images from various applications using the proposed encoder and proxy tasks. This results in a powerful pre-trained model with robust feature representation that could be utilized for various medical image analysis downstream tasks.
- We validate the effectiveness of proposed framework by fine-tuning the pre-trained Swin UNETR on two public benchmarks of MSD and BTCV and achieve *state-of-the-art* on the test leaderboards of both datasets.

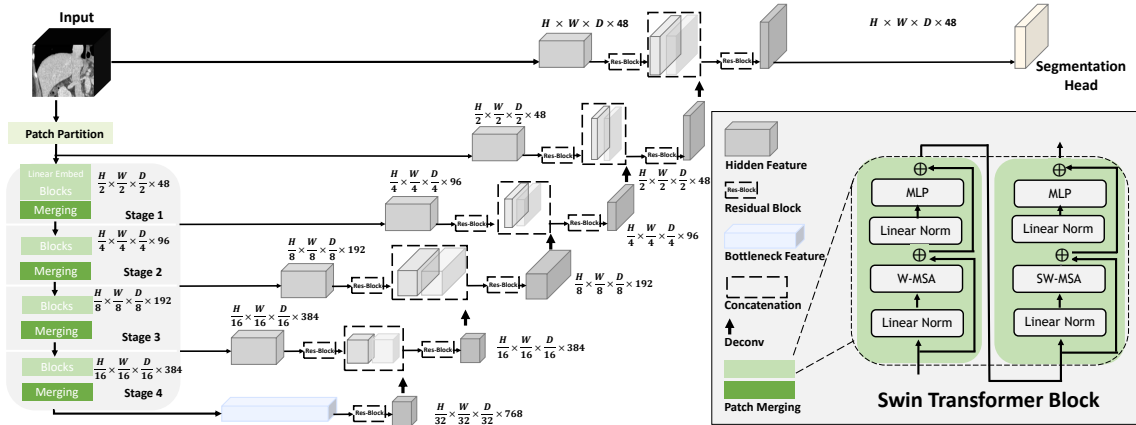


Figure 8.2: Overview of the Swin UNETR architecture.

8.2 Related Works

8.2.1 Medical Segmentation with Transformers

Vision transformers are first used in classification tasks and are adopted from sequence-to-sequence modeling in natural language processing. Self-attention mechanisms that aggregate information from the entire input sequence are first achieving comparable, then better performance against prior arts of convolutional architectures such as ResNet [223] or U-Net [115]. Recently, transformer-based networks [93, 224, 225, 226] are proposed for medical image segmentation. In these pioneering works, the transformer blocks are used either as a bottleneck feature encoder or as additional modules after convolutional layers, resulting in limited exploitation of the spatial context advantages of transformers. Comparing to prior works [93, 94], which are using transformers as secondary encoder, we propose to utilize transformers to embed high-dimensional volumetric medical images, which allow for a more direct encoding of 3D patches and positional embeddings.

Most medical image analysis tasks such as segmentation requires dense inference from multi-scale features. Skip connection-based architectures such as UNet [115] and pyramid networks [42] are widely adopted to leverage hierarchical features. However, vision transformers with a single patch size, while successful in natural image applications, are intractable for high-resolution and high-dimensional volumetric images. To avoid quadratic overflow of computing self-attention at scales [227, 228], Swin Transformer [204, 229] is proposed to construct hierarchical encoding by a shifted-window mechanism. Recent works such as Swin UNet [230] and DS-TransUNet [231] utilize the merits of Swin Transformers for 2D segmentation and achieve promising performance.

Augmenting the above-mentioned methods, we learn from 3D anatomy in broader medical image segmentation scenarios by incorporating hierarchically volumetric context.

8.2.2 Pre-training in Medical Image Analysis

In medical image analysis, previous studies of pre-training on labeled data demonstrate improved performance by transfer learning [219, 220]. However, generating annotation for medical images is expensive and time-consuming. Recent advances in self-supervised learning offer the promise of utilizing unlabeled data. Self-supervised representation learning [232, 233, 234] constructs feature embedding spaces by designing pre-text tasks, such as solving jigsaw puzzles [75]. Another commonly used pre-text task is to memorize spatial context from medical images, which is motivated by image restoration. This idea is generalized to inpainting tasks [1, 2, 74] to learn visual representations [235, 236, 237] by predicting the original image patches. Similar efforts for reconstructing spatial context have been formulated as solving Rubik’s cube problem [238], random rotation prediction [222, 239] and contrastive coding [126, 221]. Different from these efforts, our pre-training framework is simultaneously trained with a combination of pre-text tasks, tailored for 3D medical imaging data, and leverages a transformer-based encoder as a powerful feature extractor.

8.3 Swin UNETR

Swin UNETR comprises a Swin Transformer [204] encoder that directly utilizes 3D patches and is connected to a CNN-based decoder via skip connections at different resolutions. Fig. 8.2 illustrates the overall architecture of Swin UNETR. We describe the details of encoder and decoder in this section.

8.3.1 Swin Transformer Encoder

Assuming that the input to the encoder is a sub-volume $\mathcal{X} \in \mathbb{R}^{H \times W \times D \times S}$, a 3D token with a patch resolution of (H', W', D') has a dimension of $H' \times W' \times D' \times S$. The patch partitioning layer creates a sequence of 3D tokens with size $\frac{H}{H'} \times \frac{W}{W'} \times \frac{D}{D'}$ that are projected into a C -dimensional space via an embedding layer. Following [204], for efficient modeling of token interactions, we partition the input volumes into non-overlapping windows and compute local self-attention within each region. Specifically, at layer l , we use a window of size $M \times M \times M$ to evenly divide a 3D token into

$\lceil \frac{H'}{M} \rceil \times \lceil \frac{W'}{M} \rceil \times \lceil \frac{D'}{M} \rceil$ windows. In the subsequent layer $l + 1$, we shift the partitioned windows by $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$ voxels. The shifted windowing mechanism is illustrated in Fig. 8.3. The outputs of encoder blocks in layers l and $l + 1$ are computed as in

$$\begin{aligned}
\hat{z}^l &= \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1} \\
z^l &= \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l \\
\hat{z}^{l+1} &= \text{SW-MSA}(\text{LN}(z^l)) + z^l \\
z^{l+1} &= \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1},
\end{aligned} \tag{8.1}$$

where W-MSA and SW-MSA denote regular and window partitioning multi-head self-attention modules, respectively, \hat{z}^l and z^l are the outputs of W-MSA and SW-MSA; LN and MLP denote layer normalization and Multi-Layer Perceptron (see Fig. 8.2). Following [204], we adopt a 3D cyclic-shifting for efficient batch computation of shifted windowing. Furthermore, we calculate the self-attention according to

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \tag{8.2}$$

where Q, K, V represent queries, keys and values respectively, d is the size of the query and key.

Our encoder uses a patch size of $2 \times 2 \times 2$ with a feature dimension of $2 \times 2 \times 2 \times 1 = 8$ (*i.e.* single input channel CT images) and a $C = 48$ -dimensional embedding space. Furthermore, the overall architecture of the encoder consists of 4 stages comprising of 2 transformer blocks at each stage (*i.e.* $L = 8$ total layers). In between every stage, a patch merging layer is used to reduce the resolution by a factor of 2. Stage 1 consists of a linear embedding layer and transformer blocks that maintain the number of tokens as $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$. Furthermore, a patch merging layer groups patches with resolution $2 \times 2 \times 2$ and concatenates them, resulting in a $4C$ -dimensional feature embedding. A linear layer is then used to downsample the resolution by reducing the dimension to $2C$. The same procedure continues in stage 2, stage 3 and stage 4 with resolutions of $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$ and $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$ respectively. The hierarchical representations of the encoder at different stages are used in downstream applications such as segmentation for multi-scale feature extraction.

8.3.2 Decoder

The encoder of Swin UNETR is connected to a CNN-based decoder at each resolution via skip connections to create a “U-shaped” network for downstream applications such as segmentation. Specifically, we extract the output sequence representations of each stage i ($i \in \{0, 1, 2, 3, 4\}$) in the encoder as well as the bottleneck ($i = 5$) and reshape them into features with size $\frac{H}{2^i} \times \frac{W}{2^i} \times \frac{D}{2^i}$. The extracted representations at each stage are then fed into a residual block consisting of two post-normalized $3 \times 3 \times 3$ convolutional layers with instance normalization [240]. The processed features from each stage are then upsampled by using a deconvolutional layer and concatenated with processed features of the preceding stage. The concatenated features are fed into a residual block with aforementioned descriptions. For segmentation, we concatenate the output of the encoder (*i.e.* Swin Transformer) with processed features of the input volume and feed them into a residual block followed by a final $1 \times 1 \times 1$ convolutional layer with a proper activation function (*i.e.* softmax) for computing the segmentation probabilities (see Fig. 8.2 for details of the architecture).

8.4 Pre-training

We pre-train the Swin UNETR encoder with multiple proxy tasks and formulate it with a multi-objective loss function (Fig. 8.1). The objective of self-supervised representation learning is to encode region of interests (ROI)-aware information of the human body. Inspired by previous works on context reconstruction [1, 2] and contrastive encoding [241], we exploit three proxy tasks for medical image representation learning. Three additional projection heads are attached to the encoder during pre-training. Furthermore, the downstream task, e.g. segmentation, fine-tunes the full Swin UNETR model with the projection heads removed. In training, sub-volumes are cropped random regions of the volumetric data. Then, stochastic data augmentations with random rotation and cutout are applied twice to each sub-volume within a mini-batch, resulting in two views of each data.

8.4.1 Masked Volume Inpainting

The cutout augmentation masks out ROIs in the sub-volume $\mathcal{X} \in \mathbb{R}^{H \times W \times D \times C}$ randomly with volume ratio of s . We attach a transpose convolution layer to the encoder as the reconstruction head and denote its output as $\hat{\mathcal{X}}^{\mathcal{M}}$. The reconstruction objective is defined by an $L1$ loss between \mathcal{X}

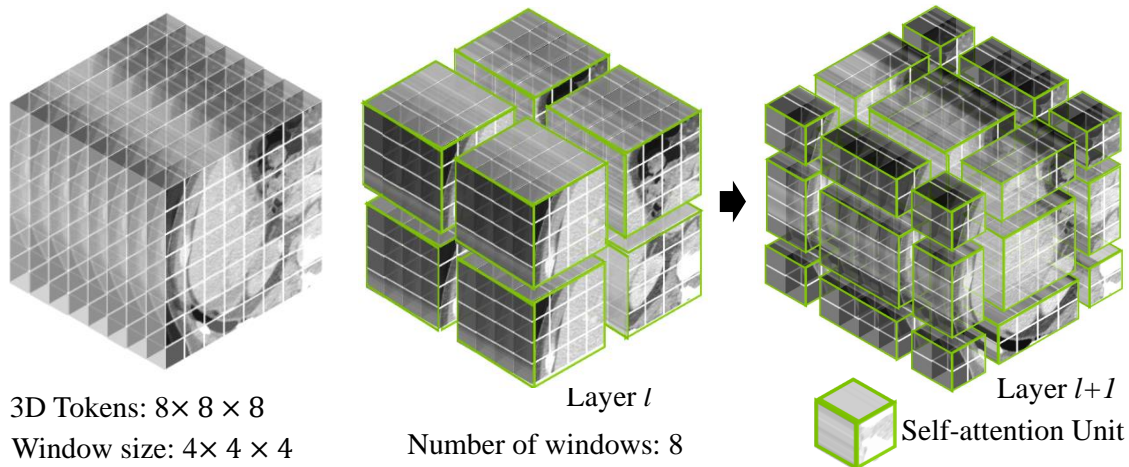


Figure 8.3: Shifted windowing mechanism for efficient self-attention computation of 3D tokens with $8 \times 8 \times 8$ tokens and $4 \times 4 \times 4$ window size.

and $\hat{\mathcal{X}}^{\mathcal{M}}$

$$\mathcal{L}_{\text{inpaint}} = \|\mathcal{X} - \hat{\mathcal{X}}^{\mathcal{M}}\|_1, \quad (8.3)$$

The masked volume inpainting is motivated by prior work which focused on 2D images [74]. We extend it to 3D domain to showcase its effectiveness on representation learning of volumetric medical images.

8.4.2 Image Rotation

The rotation prediction task predicts the angle categories by which the input sub-volume is rotated. For simplicity, we employ R classes of 0° , 90° , 180° , 270° rotations along the z -axis. An MLP classification head is used for predicting the softmax probabilities \hat{y}_r of rotation categories. Given the ground truth y_r , a cross-entropy loss is used for rotation prediction task:

$$\mathcal{L}_{\text{rot}} = - \sum_{r=1}^R y_r \log(\hat{y}_r), \quad (8.4)$$

The 3D rotation and cutout also serves simultaneously as an augmentation transformation for contrastive learning.

8.4.3 Contrastive Coding

The self-supervised contrastive coding presents promising performance on visual representation learning when transferred to downstream tasks [83, 242]. Given a batch of augmented sub-volumes, the contrastive coding allows for a better representation learning by maximizing the mutual information between positive pairs (augmented samples from same sub-volume), while minimizing that between negative pairs (views from different sub-volumes). The contrastive coding is obtained by attaching a linear layer to the Swin UNETR encoder, which maps each augmented sub-volume to a latent representation v . We use cosine similarity as the distance measurement of the encoded representations as defined in [83]. Formally, the 3D contrastive coding loss between a pair v_i and v_j is defined as:

$$\mathcal{L}_{contrast} = -\log \frac{\exp(\text{sim}(v_i, v_j)/t)}{\sum_k^{2N} \mathbf{1}_{k \neq i} \exp(\text{sim}(v_i, v_k)/t)}, \quad (8.5)$$

where t is the measurement of normalized temperature scale. $\mathbf{1}$ is the indicator function evaluating to 1 iff $k \neq i$. sim denotes the dot product between normalized embeddings. The contrastive learning loss function strengthens the intra-class compactness as well as the inter-class separability.

8.4.4 Loss Function

Formally, we minimize the total loss function by training Swin UNETR’s encoder with multiple pre-training objectives of masked volume inpainting, 3D image rotation & contrastive coding as follows:

$$\mathcal{L}_{tot} = \lambda_1 \mathcal{L}_{inpaint} + \lambda_2 \mathcal{L}_{contrast} + \lambda_3 \mathcal{L}_{rot}. \quad (8.6)$$

A grid-search hyper-parameter optimization was performed which estimated the optimal values of $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

8.5 Experiments

8.5.1 Datasets

Pre-training Datasets : A total of 5 public CT datasets, consisting of 5,050 subjects, are used to construct our pre-training dataset. The corresponding number of 3D volumes for chest, abdomen and head/neck are 2,018, 1,520 and 1,223 respectively. The collection and source details are presented in the supplementary materials. Existing annotations or labels are *not* utilized from these

Methods	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	Avg.
SETR NUP [90]	0.931	0.890	0.897	0.652	0.760	0.952	0.809	0.867	0.745	0.717	0.719	0.620	0.796
SETR PUP [90]	0.929	0.893	0.892	0.649	0.764	0.954	0.822	0.869	0.742	0.715	0.714	0.618	0.797
SETR MLA [212]	0.930	0.889	0.894	0.650	0.762	0.953	0.819	0.872	0.739	0.720	0.716	0.614	0.796
ASPP [243]	0.935	0.892	0.914	0.689	0.760	0.953	0.812	0.918	0.807	0.695	0.720	0.629	0.811
TransUNet [94]	0.952	0.927	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
CoTr* [93]	0.943	0.924	0.929	0.687	0.762	0.962	0.894	0.914	0.838	0.796	0.783	0.647	0.841
CoTr [93]	0.958	0.921	0.936	0.700	0.764	0.963	0.854	0.920	0.838	0.787	0.775	0.694	0.844
RandomPatch [43]	0.963	0.912	0.921	0.749	0.760	0.962	0.870	0.889	0.846	0.786	0.762	0.712	0.844
PaNN [44]	0.966	0.927	0.952	0.732	0.791	0.973	0.891	0.914	0.850	0.805	0.802	0.652	0.854
nnUNet [58]	0.967	0.924	0.957	0.814	0.832	0.975	0.925	0.928	0.870	0.832	0.849	0.784	0.888
UNETR [214]	0.972	0.942	0.954	0.825	0.864	0.983	0.945	0.948	0.890	0.858	0.852	0.812	0.891
Swin UNETR	0.976	0.958	0.956	0.893	0.875	0.985	0.953	0.949	0.904	0.899	0.898	0.846	0.918

Table 8.1: Leaderboard¹Dice results of BTCV challenge on multi-organ segmentation. The proposed method achieves state-of-the-art performance in both free and standard competitions. Note: Spl: spleen, RKid: right kidney, LKid: left kidney, Gall: gallbladder, Eso: esophagus, Liv: liver, Sto: stomach, Aor: aorta, IVC: inferior vena cava, Veins: portal and splenic veins, Pan: pancreas, AG: left and right adrenal glands.

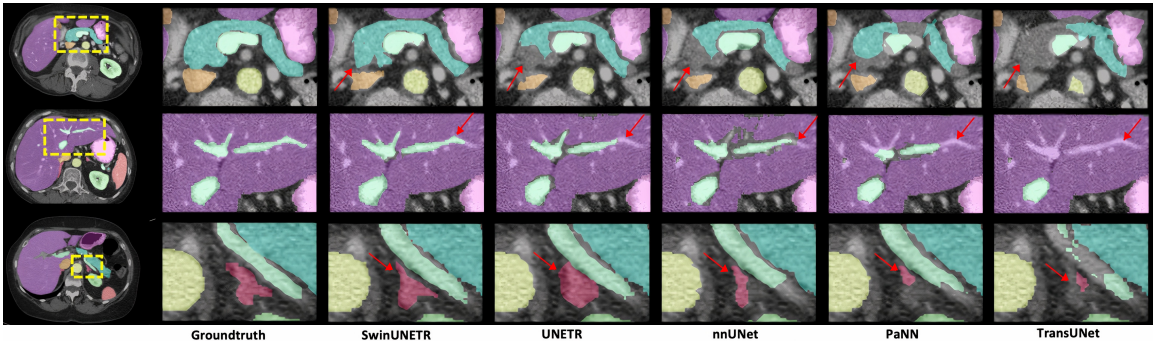


Figure 8.4: Qualitative visualizations of the proposed Swin UNETR and baseline methods. Three representative subjects are demonstrated. Regions of evident improvements are enlarged to show better details of pancreas (blue), portal vein (light green), and adrenal gland (red).

datasets during the pre-training stage.

BTCV : The Beyond the Cranial Vault (BTCV) abdomen challenge dataset [72] contains 30 subjects with abdominal CT scans where 13 organs are annotated by interpreters under supervision of radiologists at Vanderbilt University Medical Center. Each CT scan is acquired with contrast enhancement phase at portal venous consists of 80 to 225 slices with 512×512 pixels and slice thickness ranging from 1 to 6 mm. The multi-organ segmentation problem is formulated as a 13 classes segmentation task (see Table 8.1 for details). The pre-processing pipeline is detailed in supplementary materials.

MSD: Medical Segmentation Decathlon (MSD) dataset [211] comprises of 10 segmentation tasks from different organs and image modalities. These tasks are designed to feature difficulties

¹<https://www.synapse.org/#!/Synapse:syn3193805/wiki/217785/>

across medical images, such as small training sets, unbalanced classes, multi-modality data and small objects. Therefore, the MSD challenge can serve as a comprehensive benchmark to evaluate the generalizability of medical image segmentation methods. The pre-processing pipeline for this dataset is outlined in supplementary materials.

8.5.2 Implementation Details

For pre-training tasks, (1) masked volume inpainting: the ROI dropping rate is set to 30% (as also used in [232]); the dropped regions are randomly generated and they sum up to reach overall number of voxels; (2) 3D contrastive coding: a feature size of 512 is used as the embedding size; (3) rotation prediction: the rotation degree is configured to 0° , 90° , 180° , and 270° . We train the model using the AdamW [244] optimizer with a warm-up cosine scheduler of 500 iterations. The pre-training experiments use a batch-size of 4 per GPU (with $96 \times 96 \times 96$ patch), and initial learning rate of $4e^{-4}$, momentum of 0.9 and decay of $1e^{-5}$ for 450K iterations. Our model is implemented in PyTorch and MONAI². A five-fold cross validation strategy is used to train models for all BTCV and MSD experiments. We select the best model in each fold and ensemble their outputs for final segmentation predictions. Detailed training hyperparameters for fine-tuning BTCV and MSD tasks can be found in the supplementary materials. All models are trained on a NVIDIA DGX-1 server.

8.5.3 Evaluation Metrics

The Dice similarity coefficient (Dice) and Hausdorff Distance 95% (HD95) are used as measurements for experiment results. HD95 calculates 95th percentile of surface distances between ground truth and prediction point sets. Metric formulations are as follows:

$$\text{Dice} = \frac{2 \sum_{i=1}^I Y_i \hat{Y}_i}{\sum_{i=1}^I Y_i + \sum_{i=1}^I \hat{Y}_i}, \quad (8.7)$$

$$\text{HD} = \max\left\{\max_{y' \in Y'} \min_{\bar{y}' \in \bar{Y}'} \|y' - \bar{y}'\|, \max_{\bar{y}' \in \bar{Y}'} \min_{y' \in Y'} \|\bar{y}' - y'\|\right\}. \quad (8.8)$$

where Y and \bar{Y} denote the ground truth and prediction of voxel values. Y' and \bar{Y}' denote ground truth and prediction surface point sets. Surface Dice [245] is also used, which is referred as Normalized Surface Distance (NSD) in MSD challenge evaluation. The metric measures the overlap of ground

²<https://monai.io/>

Method	Rank	Average Accuracy	
		Dice \uparrow	NSD \uparrow
Swin UNETR	1	78.68	89.28
DiNTS [246]	2	77.93	88.68
nnUNet [58]	3	77.89	88.09
Models Gen. [2]	4	76.97	87.19
Trans VW [1]	5	76.96	87.64

Table 8.2: Overall performance of top-ranking methods on all 10 segmentation tasks in the MSD public test leaderboard. NSD denotes Normalized Surface Distance.

Organ	Task01 Brain Tumour								Task03 Liver				Task06 Lung			
Metric	Dice1	Dice2	Dice3	Avg.	NSD1	NSD2	NSD3	Avg.	Dice1	Dice2	Avg.	NSD1	NSD2	Avg.	Dice1	NSD1
Kim et al [247]	67.40	45.75	68.26	60.47	86.65	72.03	90.28	82.99	94.25	72.96	83.61	96.76	88.58	92.67	63.10	62.51
Trans VW [1]	68.03	46.98	68.40	61.14	87.52	72.42	90.91	83.62	95.18	76.90	86.04	97.86	92.03	94.95	74.54	76.22
C2FNAS[248]	67.62	48.60	69.72	61.98	87.61	72.87	91.16	83.88	94.98	72.89	83.94	98.38	89.15	93.77	70.44	72.22
Models Gen. [2]	68.03	46.98	68.40	61.14	87.52	72.42	90.91	83.62	95.72	77.50	86.61	98.48	91.92	95.20	74.54	76.22
nnUNet [58]	68.04	46.81	68.46	61.10	87.51	72.47	90.78	83.59	95.75	75.97	85.86	98.55	90.65	94.60	73.97	76.02
DiNTS [246]	69.28	48.65	69.75	62.56	89.33	73.16	91.69	84.73	95.35	74.62	84.99	98.69	91.02	94.86	74.75	77.02
Swin UNETR	70.02	52.52	70.51	64.35	89.07	80.30	93.46	87.61	95.35	75.68	85.52	98.34	91.59	94.97	76.60	77.40

Table 8.3: MSD test dataset performance comparison of Dice and NSD. Benchmarks obtained from MSD test leaderboard

truth and prediction surfaces (with a fixed tolerance) instead of the overlap of two volumes. This provides a measure of agreement between the surfaces of two structures.

8.5.4 Results

8.5.4.1 BTCV Multi-organ Segmentation Challenge

We extensively compare the benchmarks of our model with baselines. The published leaderboard evaluation is shown in Table 8.2. Compared with other top submissions, the proposed Swin UNETR achieves the best performance. We obtain the state-of-the-art Dice of 0.908, outperforming the second, third and fourth top-ranked baselines by 1.6%, 2.0% and 2.4% on average of 13 organs, respectively. Distinct improvements can be specifically observed for organs that are smaller in size,

³<https://decathlon-10.grand-challenge.org/evaluation/challenge/leaderboard/>

Organ	Task07 Pancreas						Task08 Hepatic Vessel						Task09 Spleen		Task10 Colon	
Metric	Dice1	Dice2	Avg.	NSD1	NSD2	Avg.	Dice1	Dice2	Avg.	NSD1	NSD2	Avg.	Dice1	NSD1	Dice1	NSD1
Kim et al [247]	80.61	51.75	66.18	95.83	73.09	84.46	62.34	68.63	65.49	83.22	78.43	80.83	91.92	94.83	49.32	62.21
Trans VW [1]	81.42	51.08	66.25	96.07	70.13	83.10	65.80	71.44	68.62	84.01	80.15	82.08	97.35	99.87	51.47	60.53
C2FNAS[248]	80.76	54.41	67.59	96.16	75.58	85.87	64.30	71.00	67.65	83.78	80.66	82.22	96.28	97.66	58.90	72.56
Models Gen. [2]	81.36	50.36	65.86	96.16	70.02	83.09	65.80	71.44	68.62	84.01	80.15	82.08	97.35	99.87	51.47	60.53
nnUNet [58]	81.64	52.78	67.21	96.14	71.47	83.81	66.46	71.78	69.12	84.43	80.72	82.58	97.43	99.89	58.33	68.43
DiNTS [246]	81.02	55.35	68.19	96.26	75.90	86.08	64.50	71.76	68.13	83.98	81.03	82.51	96.98	99.83	59.21	70.34
Swin UNETR	81.85	58.21	70.71	96.57	79.10	87.84	65.69	72.20	68.95	84.83	81.62	83.23	96.99	99.84	59.45	70.89

Table 8.4: MSD test dataset performance comparison of Dice and NSD. Benchmarks obtained from MSD test leaderboard³.

such as splenic and portal veins of 3.6% against prior state-of-the-art method, pancreas of 1.6%, and adrenal glands of 3.8%. Moderate improvements are observed in other organs. The representative samples in Fig. 8.4 demonstrate the success of identifying organ details by Swin UNETR. Our method detects the pancreas tail (row 1), and branches in the portal vein (row 2) in Fig. 8.4, where other methods under segment parts of each tissue. In addition, our method demonstrates distinct improvement in segmentation of adrenal glands (row 3).

8.5.4.2 Segmentation Results on MSD

The overall MSD results per task and ranking from the challenge leaderboard are shown in Table 8.3 and 8.4. The proposed Swin UNETR achieves state-of-the-art performance in Task01 BrainTumour, Task06 Lung, Task07 Pancreas, and Task10 Colon. The results are comparable for Task02 Heart, Task03 Liver, Task04 Hippocampus, Task05 Prostate, Task08 HepaticVessel and Task09 Spleen. Overall, Swin UNETR presents the best average Dice of 78.68% across all ten tasks and achieves the top ranking in the MSD leaderboard. The detail number of multiple tasks are shown in Table 8.3. Qualitative visualization can be observed in Fig. 8.5. Swin UNETR with self-supervised pre-training demonstrates visually better segmentation results in the CT tasks. The pre-trained weights are only used for fine-tuning CT tasks including Liver, Lung, Pancreas, HepaticVessel, Spleen, and Colon. For MRI tasks: Brain Tumour, Heart, Hippocampus, Prostate, experiments are trained from scratch because of the domain gap between CT and MRI images. Due to space limitations, we present the MSD test benchmarks for the remaining three MRI tasks in the supplementary materials.

8.5.5 Ablation Study

8.5.5.1 Efficacy of Pre-training

A comparison of all MSD CT tasks using pre-trained model against training from scratch can be observed in Fig. 8.6. Distinct improvement can be observed for Task03 Liver, Dice of 77.77% comparing to 75.27%. Task08 Hepatic Vessel achieves 68.52% against 64.63%. Task10 Colon shows the largest improvement, from 34.83% to 43.38%. Task07 Pancreas and Task09 Spleen both achieve significant improvement from 67.12% to 67.82%, and 96.05% to 97.32% respectively.

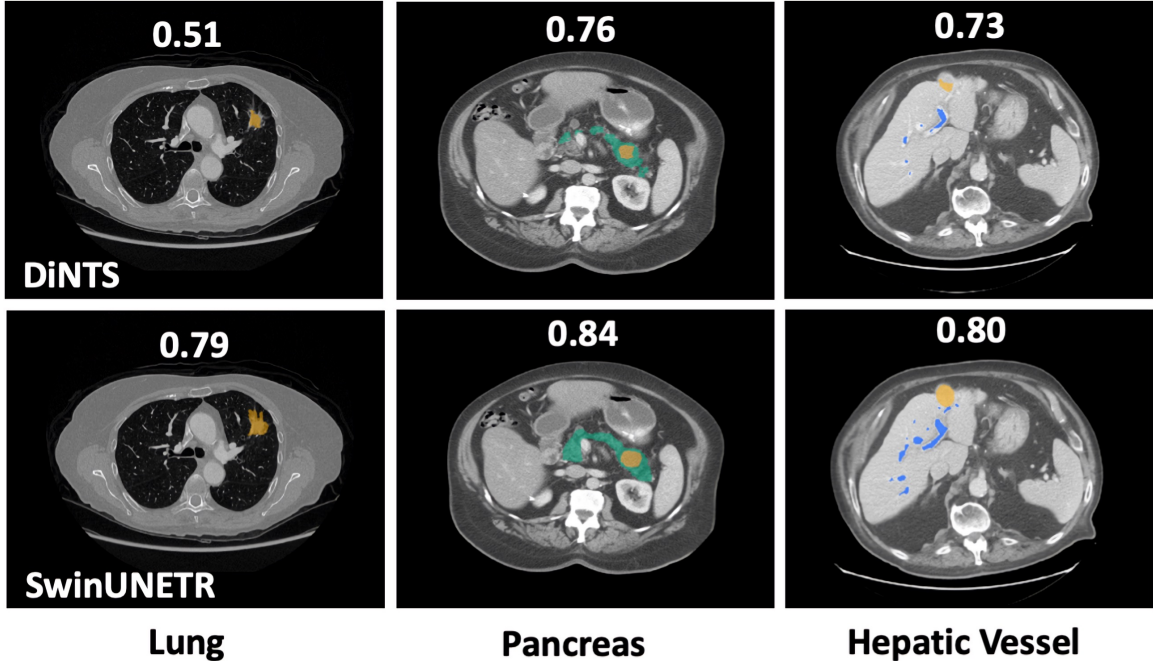


Figure 8.5: Qualitative results of representative MSD CT tasks. Average Dice values are illustrated on top of each image. Our model demonstrates more accurate performance in comparison to DiNTS for both organ and tumor segmentation across different tasks.

Loss Function	Average Accuracy	
	Dice \uparrow	HD \downarrow
Scratch	83.43	42.36
\mathcal{L}_{rot}	83.56	36.19
$\mathcal{L}_{contrast}$	83.67	38.81
$\mathcal{L}_{inpaint}$	83.85	28.94
$\mathcal{L}_{inpaint} + \mathcal{L}_{rot}$	84.01	26.06
$\mathcal{L}_{inpaint} + \mathcal{L}_{contrast}$	84.45	24.37
$\mathcal{L}_{inpaint} + \mathcal{L}_{contrast} + \mathcal{L}_{rot}$	84.72	20.03

Table 8.5: Ablation study of the effectiveness of each objective function in the proposed pre-training loss. HD denotes Hausdorff Distance. Experiments on fine-tuning the BTCV dataset.

8.5.5.2 Reduce Manual Labeling Efforts

Fig. 8.7 demonstrates the comparison results of fine-tuning using a subset of BTCV dataset. We show using 10% of labeled data, experiments with pre-training weights achieve approximately 10% improvement comparing to training from scratch. On employing all labeled data, the self-supervised pre-training shows 1.3% higher average Dice. The Dice number 83.13 of learning from scratch with entire dataset can be achieved by using pre-trained Swin UNETR with 60% data. Fig. 8.7 indicates

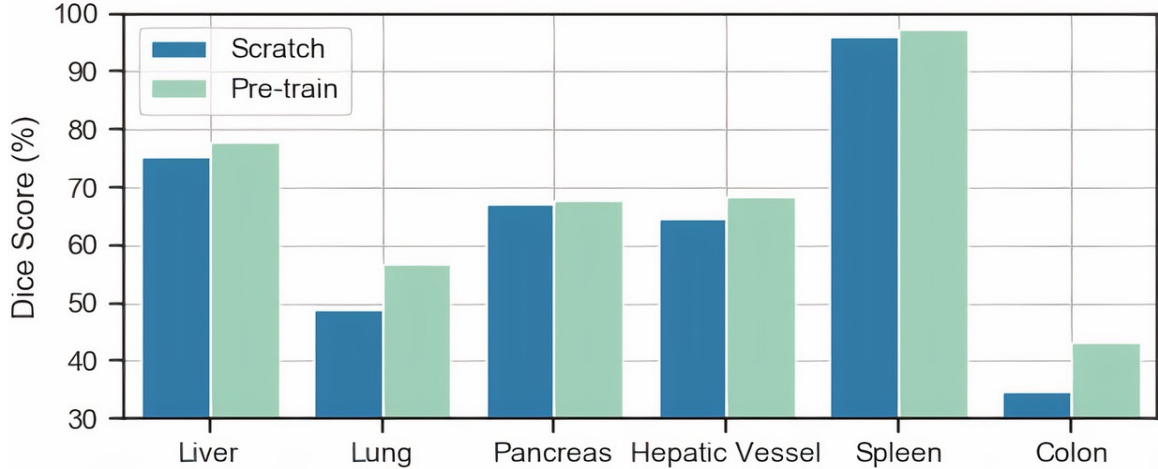


Figure 8.6: The indication of Dice gap between using pre-training (Green) and scratch model (Blue) on MSD CT tasks validation set.

that our approach can reduce the annotation effort by at least 40% for BTCV task.

8.5.5.3 Size of Pre-training Dataset

We perform organ-wise study on BTCV dataset by using pre-trained weights of smaller unlabeled data. In Fig. 8.8, the fine-tuning results are obtained from pre-training 100, 3,000, and 5,000 scans. We observe that Swin UNETR is robust with respect to the total number of CT scans trained. Fig. 8.8 demonstrates the proposed model can benefit from larger pre-training datasets with increasing size of unlabeled data.

8.5.5.4 Efficacy of Self-Supervised Objectives

We perform empirical study on pre-training with different combinations of self-supervised objectives. As shown in Table 8.5, on BTCV test set, using pre-trained weights by inpainting achieves the highest improvement at single task modeling. On pairing tasks, inpainting and contrastive learning show Dice of 84.45% and Hausdorff Distance (HD) of 24.37. Overall, employing all proxy tasks achieves best Dice of 84.72%.

8.6 Discussion and Limitations

Our state-of-the-art results on the test leaderboards of MSD and BTCV datasets validate the effectiveness of the proposed self-supervised learning framework in taking the advantage of large number

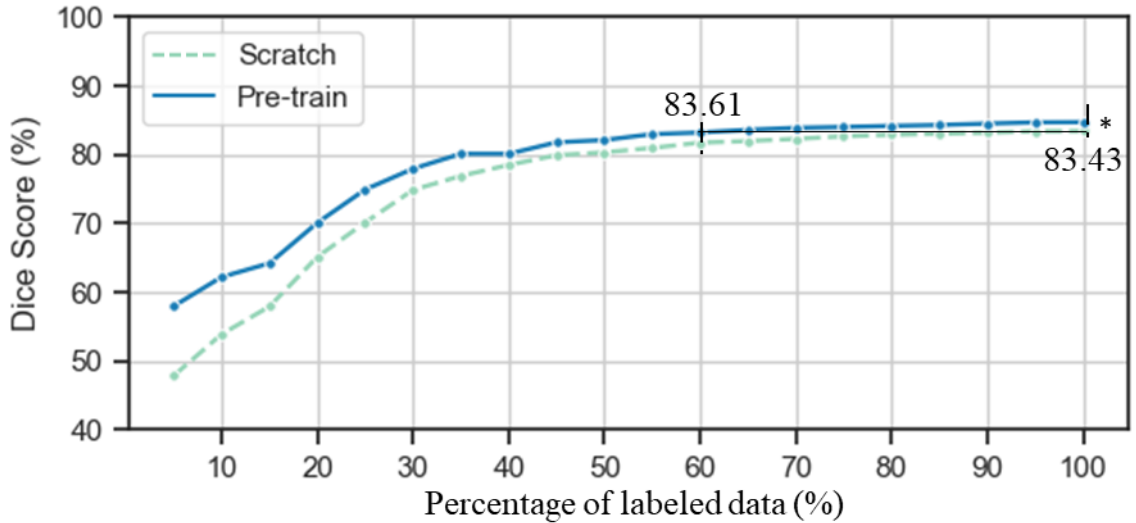


Figure 8.7: Data-efficient performance on BTCV test dataset. Significance under Wilcoxon Signed Rank test, * : $p < 0.001$.

of available medical images without the need of annotation effort. Subsequently, fine-tuning the pre-trained Swin UNETR model achieves higher accuracy, improves the convergence speed, and reduces the annotation effort in comparison to training with randomly initialized weights from scratch. Our framework is scalable and can be easily extended with more proxy tasks and augmentation transformations. Meanwhile, the pre-trained encoder can benefit the transfer learning of various medical imaging analysis tasks, such as classification and detection. In MSD pancreas segmentation task, Swin UNETR with pre-trained weights outperforms AutoML algorithms such as DiNTS [246] and C2FNAS [248] that are specifically designed for searching the optimal network architectures on the same segmentation task. Currently, Swin UNETR has only been pre-trained using CT images, and our experiments have not demonstrated enough transferability when applied directly to other medical imaging modalities such as MRI. This is mainly due to obvious domain gaps and different number of input channels that are specific to each modality. As a result, this is a potential direction that should be studied in future efforts.

8.7 Conclusions

In this work, we present a novel framework for self-supervised pre-training of 3D medical images. Inspired by merging feature maps at scales, we built the Swin UNETR by exploiting transformer-encoded spatial representations into convolution-based decoders. By proposing the first transformer-

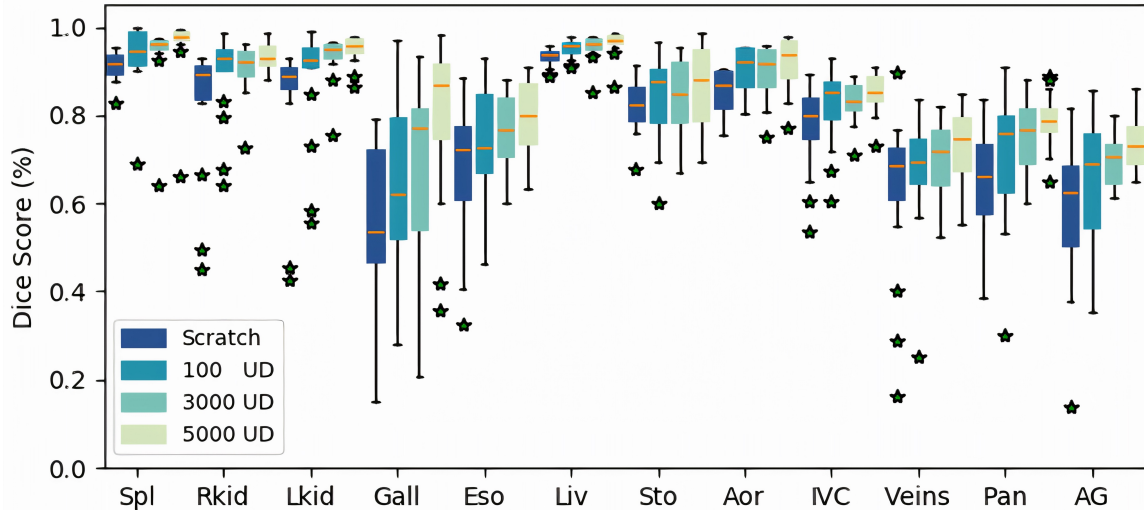


Figure 8.8: Pre-trained weights using 100, 3000 and 5000 scans are compared for fine-tuning on the BTCV dataset for each organ.

based 3D medical image pre-training, we leverage the power of Swin Transformer encoder for fine-tuning segmentation tasks. Swin UNETR with self-supervised pre-training achieves the state-of-the-art performance on the BTCV multi-organ segmentation challenge and MSD challenge. Particularly, we presented the large-scale CT pre-training with 5,050 volumes, by combining multiple publicly available datasets and diversities of anatomical ROIs.

8.8 Appendix

We provide the supplementary materials in the following. In Sec. 8.9, we describe the details of datasets that are used for pre-training from public sources. In Sec. 8.10, we illustrate the preprocessing and implementation details of fine-tuning tasks using BTCV and MSD datasets. In Sec. 8.11, we present qualitative and quantitative comparisons of segmentation tasks in MRI modality from MSD dataset. The presented results include benchmarks from all top-ranking methods using the MSD test leaderboard. In Sec. 8.12, the model complexity analysis is presented. Finally, we provide pseudocode of Swin UNETR self-supervised pre-training in Sec. 8.13.

8.9 Pre-training Datasets

In this section, we provide additional information for our pre-training datasets. The proposed Swin UNETR is pre-trained using five collected datasets. The total data cohort contains 5,050

CT scans of various body region of interests (ROI) such as head, neck, chest, abdomen, and pelvis. LUNA16 [249], TCIA Covid19 [250] and LiDC [251] contain 888, 761 and 475 CT scans which composes the chest CT cohort. The HNSCC [252] has 1,287 CT scans from head and neck squamous cell carcinoma patients. The TCIA Colon dataset [253] comprises the abdomen and pelvis cohort with 1,599 scans. We split 5% of each dataset for validation in the pre-training stage. Table 8.6 summarizes sources of each collected dataset. Overall, the number of training and validation volumes are 4,761 and 249, respectively. The Swin UNETR encoder is pre-trained using only unlabeled images, annotations were not utilized from any of these datasets. We first clip CT image intensities from -1000 to 1000 , then normalize to 0 and 1. To obtain informative patches of covering anatomies, we crop sub-volumes of $96 \times 96 \times 96$ voxels at foregrounds, and exclude full air (voxel = 0) patches. In summary, Swin UNETR is pre-trained via a diverse set of human body compositions, and learn a general-purpose representation from different institutes' data that can be leveraged for wide range of fine-tuning tasks.

8.10 Preprocessing Pipelines

We report fine-tuning results on two public benchmarks: BTCV [72] and MSD challenge [51]. BTCV contains 30 CT scans with 13 annotated anatomies and can be formulated as a single multi-organ segmentation task. The MSD contains 10 tasks for multiple organs, from different sources and using different modalities. Details regarding preprocessing these datasets are provided in the subsequent sub-sections of 2.1 and 2.2.

8.10.1 BTCV Dataset

All CT scans are interpolated into the isotropic voxel spacing of $[1.5 \times 1.5 \times 2.0]$ mm. The multi-organ segmentation problem is formulated as a 13 class segmentation, which includes large organs such as liver, spleen, kidneys and stomach; vascular tissues of esophagus, aorta, IVC, splenic and portal veins; small anatomies of gallbladder, pancreas and adrenal glands. Soft tissue window is used for clipping the CT intensities, then normalized to 0 and 1 followed by random sampling of $96 \times 96 \times 96$ voxels. Data augmentation of random flip, rotation and intensities shifting are used for training, with probabilities of 0.1, 0.1, and 0.5, respectively.

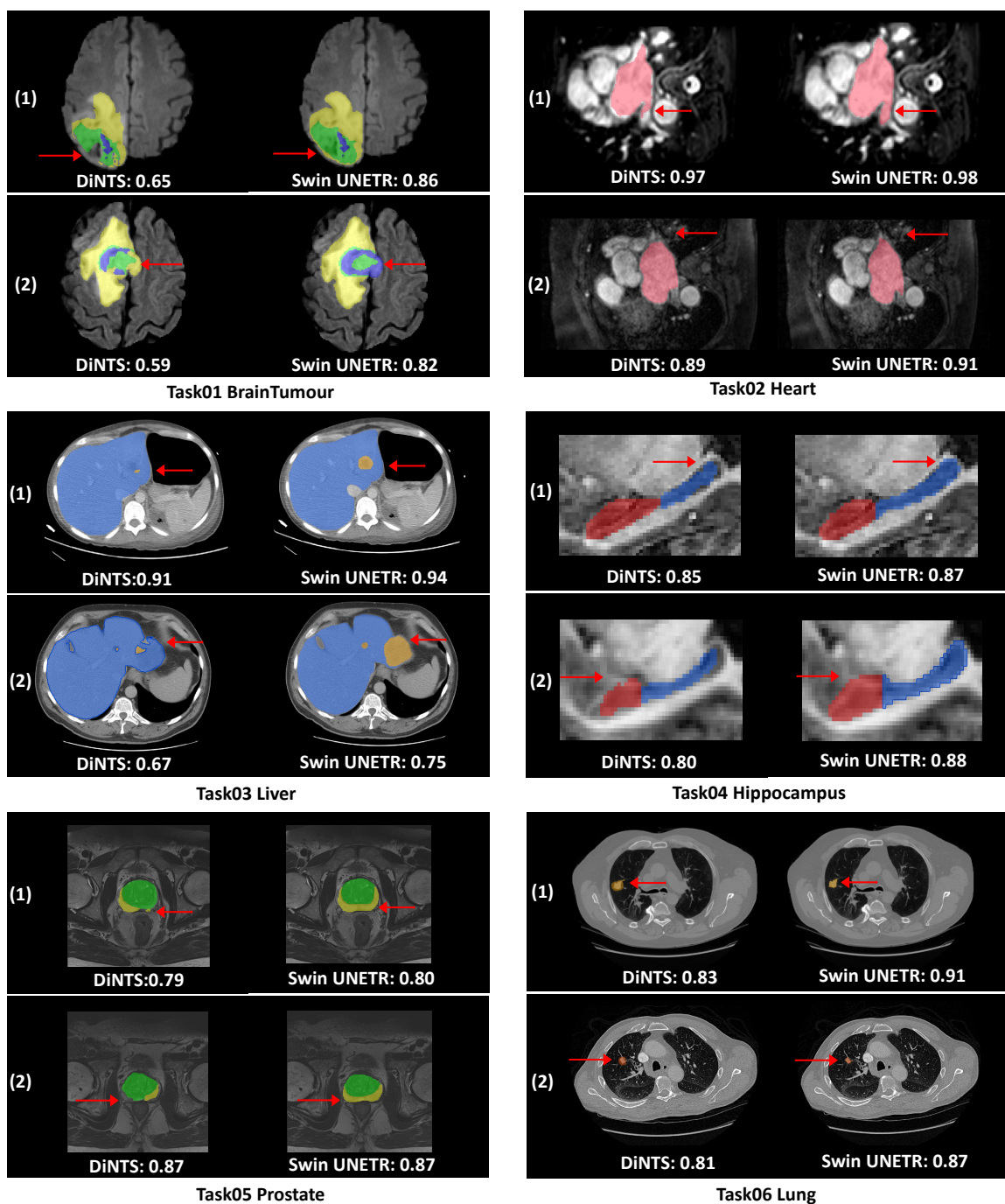


Figure 8.9

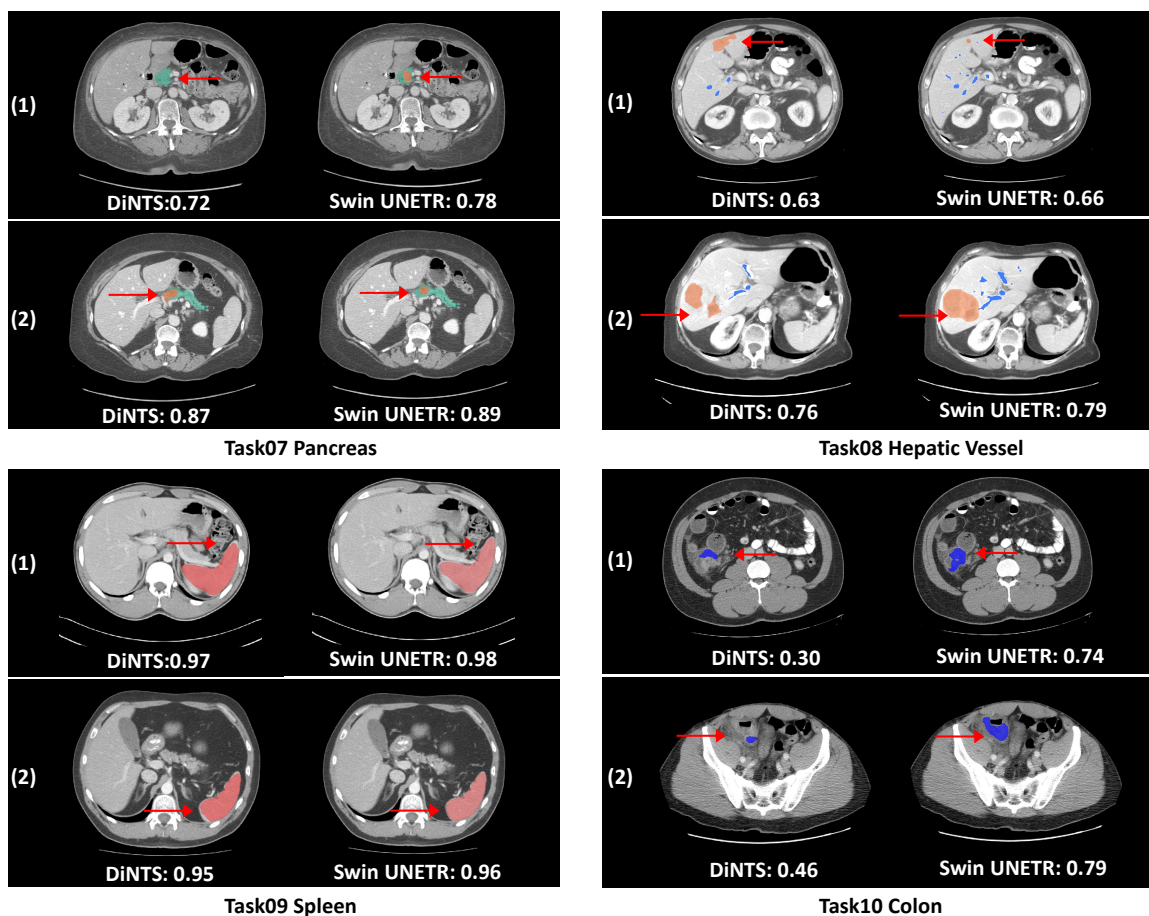


Figure 8.10: Qualitative visualizations of the proposed Swin UNETR and DiNTS on MSD Tasks

Dataset	Region of Interest	#Total Samples	Source	Train/Validation
LUNA16 [249]	Chest	888	luna16.grand-challenge.org/Data/	844/44
TCIA Covid19 [250]	Chest	761	wiki.cancerimagingarchive.net/display/Public/COVID-19	723/38
HNSCC [252]	Head/Neck	1287	wiki.cancerimagingarchive.net/display/Public/HNSCC	1223/64
TCIA Colon [253]	Abdomen/pelvis	1599	www.cancerimagingarchive.net/collections/	1520/79
LIDC [251]	Chest	475	wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI	451/24

Table 8.6: Summary of datasets for pre-training, the use of cohorts identifies diversified regions of interest.

8.10.2 MSD Dataset

The MSD challenge contains 6 CT and 4 MRI datasets. We provide additional parameters of pre-processing and augmentation details for each task as follows:

Task01 BrainTumour: The four modalities MRI images for each subject are formed into 4 channels input. We convert labels to multiple channels based on tumor classes. which label 1 is the peritumoral edema, label 2 is the GD-enhancing tumor, and label 3 is the necrotic and non-enhancing tumor core. Label 2 and 3 are merged to construct tumor core (TC), label 1, 2 and 3 are merged to

Organ	Task02 Heart		Task04 Hippocampus						Task05 Prostate						MRI tasks Avg	
Metric	DSC1	NSD1	DSC1	DSC2	Avg.	NSD1	NSD2	Avg.	DSC1	DSC2	Avg.	NSD1	NSD2	Avg.	DSC	NSD
Kim et al [247]	93.11	96.44	90.11	88.72	89.42	97.77	97.73	97.75	72.64	89.02	80.83	95.05	98.03	96.54	80.96	93.43
Trans VW [1]	93.33	96.51	90.29	88.77	89.53	97.87	97.67	97.77	73.69	88.88	81.29	95.42	98.52	96.97	81.32	93.72
C2FNAS[248]	92.49	95.81	89.37	87.96	88.67	97.27	97.35	97.31	74.88	88.75	81.82	98.79	95.12	96.96	81.24	93.49
Models Gen [2]	93.33	96.51	90.29	88.77	89.53	97.87	97.67	97.77	73.69	88.88	81.29	95.42	98.52	96.97	81.32	93.72
nnUNet [58]	93.30	96.74	90.23	88.69	89.46	97.79	97.53	97.75	76.59	89.62	83.11	96.27	98.85	97.56	81.74	93.91
DiNTS [246]	92.99	96.35	89.91	88.41	89.16	97.76	97.56	97.66	75.37	89.25	82.31	95.96	98.82	97.39	81.76	94.03
SwinUNETR	92.62	96.23	89.95	88.42	89.19	97.63	97.32	97.48	75.65	89.15	82.40	95.89	98.70	97.30	82.14	94.66

Table 8.7: Additional MSD MRI test dataset performance comparison of Dice and NSD. Benchmarks obtained from MSD test leaderboard. Task01 BrainTumour results are shown in the paper. Note: The results reported for TransVW [1] and Models Genesis [2] are from the official leaderboard for MRI tasks.

Models	#Params (M)	FLOPs (G)	Inference Time (s)
nnUNet [58]	19.07	412.65	10.28
CoTr [93]	46.51	399.21	19.21
TransUNet [94]	96.07	48.34	26.97
ASPP [180]	47.92	44.87	25.47
SETR [90]	86.03	43.49	24.86
UNETR	92.58	41.19	12.08
SwinUNETR	61.98	394.84	13.84

Table 8.8: Comparison of number of parameters, FLOPs and averaged inference time for various models in BTCV experiments.

construct whole tumor (WT), and label 2 is the enhancing tumor (ET). We crop the sub-volume of $128 \times 128 \times 128$ voxels and use channel-wise nonzero normalization for MRI images. Data augmentation probabilities of 0.5, 0.1 and 0.1 are set for random flips at each axis, intensities scaling and shifting, respectively.

Task02 Heart: The heart MRI images are interpolated to the isotropic voxel spacing of 1.0 mm. Channel-wise nonzero normalization is applied to each scan. We sample the training sub-volumes of $96 \times 96 \times 96$ voxels by ratio of positive and negative as 2:1. Augmentation probabilities for random flip, rotation, intensities scaling and shifting are set to 0.5, 0.1, 0.2, 0.5, respectively.

Task03 Liver: Each CT scan is interpolated to the isotropic voxel spacing of 1.0 mm. Intensities are scaled to $[-21, 189]$, then normalized to $[0, 1]$. 3D patches of $96 \times 96 \times 96$ voxels are obtained by sampling positive and negative ratio of 1 : 1. Data augmentation of random flip, rotation, intensities scaling and shifting are used, for which the probabilities are set to 0.2, 0.2, 0.1, 0.1, respectively.

Task04 Hippocampus: Each hippocampus MRI image is interpolated by voxel spacing of $0.2 \times 0.2 \times 0.2$, then applied spatial padding to $96 \times 96 \times 96$ as the input size of Swin UNETR model. Same as other MRI datasets, channel-wise nonzero normalization is used for intensities. Probability

of 0.1 is used for random flip, rotation, intensity scaling & shifting.

Task05 Prostate: We utilize both given modalities for prostate MRI images for each subject as two channels input. Channel-wise nonzero normalization is used. Voxel spacing of 0.5 and spatial padding of each axis are employed to construct the input size of $96 \times 96 \times 96$. We use random flip, rotation, intensity scaling and shifting with probabilities of 0.5 as data augmentations. Random affine is applied as additional transformation with scale factor of $[0.3, 0.3, 0.0]$ and rotation range of $[0, 0, \pi]$ at each axis.

Task06 Lung: We interpolate each image to isotropic voxel spacing of 1.0. Hounsfield unit (HU) range of $[-1000, 1000]$ is used and normalized to $[0, 1]$. Subsequently, training sample are cropped to $96 \times 96 \times 96$ with positive and negative ratio of 2 : 1. Augmentation probabilities of 0.5, 0.3, 0.1, 0.1 are used for random flip, rotation, intensities scaling and shifting.

Task07 Pancreas: We clip the intensities to a range of -87 to 199 . Patch size of $96 \times 96 \times 96$ is used to sample training data with positive and negative ratio of 1 : 1. We set augmentation of random flip, rotation and intensity scaling to probabilities of 0.5, 0.25 and 0.5, respectively.

Task08 HepaticVessel: To fit the optimal tissue window for hepatic vessel and tumor, we clip each CT image intensities to $[0, 230]$ HU. We apply data augmentation same with Task07 Pancreas for training.

Task09 Spleen: Spleen CT scans are pre-process with interpolation isotropic voxel spacing of 1.0 *mm* on each axis. Soft tissue window of $[-125, 275]$ HU is used for the portal venous phase contrast enhanced CT images. We use the training data augmentation of random flip, intensity scaling & shifting with probabilities of 0.15, 0.1, and 0.1, respectively.

Task10 Colon: We use HU range of $[-57, 175]$ for the colon tumor segmentation task and normalized to 0 and 1. Next, we sample training sub-volumes by positive and negative ratio of 1 : 1. Same as Task07 and Task08, we use random flip, rotation, intensity scaling as augmentation transforms with probabilities of 0.5, 0.25 and 0.5, respectively.

8.11 Results

8.11.1 MSD Qualitative Comparisons

In this section, we provide extensive segmentation visualization from MSD dataset. In particular, we compare two cases randomly selected from Swin UNETR and DiNTS for each MSD task. As

shown in Fig 8.10, DiNTS includes the under-segmentation due to lack of parts of labels (Heart, Hippocampus). The missing parts result in a lower Dice score. On BrainTumour, Liver, Pancreas, HepaticVessel and Colon tasks, the comparison indicate that our method achieves better segmentation where the under-segmentation of tumors are observed in DiNTS. For Lung task, the over-segmentation is observed with DiNTS where surrounding tissues are included with label of the lung cancer, while Swin UNETR clearly delineate the boundary. In Heart and Spleen, DiNTS and Swin UNETR have comparable Dice score, yet Swin UNETR performs better segmentation on tissue corner (See Fig 8.10). Overall, Swin UNETR achieves better segmentation results and solves the under- and over-segmentation outliers as observed in segmentation via DiNTS.

8.11.2 MSD Quantitative Comparisons

In this section, we provide the quantitative benchmarks of MRI segmentation tasks from MSD dataset. In addition to Task01 BrainTumour, we implement experiment on three remaining MRI dataset including Heart, Hippocampus and Prostate (see Table. 8.7). The results are directly obtained from the MSD⁴ leaderboard. Regarding MRI benchmark, we achieve much better performance on brain tumor segmentation presented in the paper, with average Dice improvement of 2% against second best performance. Comparing to models genesis [2], nnUNet [58], the Swin UNETR shows comparable results on Heart, Hippocampus and Prostate. Overall, we achieve the best average results (Dice of 82.14% and NSD of 94.66%) across four MRI datasets, showing Swin UNETR’s superiority of medical image segmentation.

8.12 Model Complexity and Pre-training Time

In this section, we examine the model complexity along with inference time. In Table. 8.8, the number of network paramerts, FLOPs, and averaged inference time of Swin UNETR and baselines on BTCV dataset are presented. We calculate the FLOPs and inference time based on input size of $96 \times 96 \times 96$ used in the BTCV experiments with sliding window approach. Swin UNETR shows moderate size of parameter with 61.98M, less than transformer-based methods such as TransUNet [94] of 96.07M, SETR [212] of 86.03M, and UNETR [214] of 92.58M, but larger than 3DUNet (nnUNet) [58] of 19.07M, ASPP [180] 47.92M. Our model also shows comparable FLOPs

⁴<https://decathlon-10.grand-challenge.org/evaluation/challenge/leaderboard/>

and inference time in terms of 3D approaches such as nnUNet [58] and CoTr [93]. Overall, Swin UNETR outperforms CNN-based and other transformer-based methods while preserves moderate model complexity. Regarding self-supervised pre-training time of Swin UNETR encoder, our approach takes only approximately 6 GPU days. We evaluate pre-training on the 5 collected public datasets with totally 5,050 scans for training and validation, and set maximum training iterations to 45K steps.

8.13 Pre-Training Algorithm Details

In this section, we illustrate the Swin UNETR pre-training details. The Swin UNETR is trained in self-supervised learning paradigm, where we design masked volume inpainting, rotation prediction and contrastive coding as proxy tasks. The self-training aims at improving the quality of representations learnt by large unlabeled data and propagating to smaller fine-tuning dataset. To this end, we leverage multiple transformations for input 3D data, which can exploit inherent context by a mechanism akin to autoencoding and similarity identification. In particular, given an input mini batch data, the transform of random rotation is implemented on each image in the mini batch iteratively. To simultaneously utilize augmentation transformations for contrastive learning, the random rotation of 0° , 90° , 180° , 270° is applied twice on the same input to generate randomly augmented image pairs of the same image patch. Subsequently, the mini batch data pairs are constructed with the cutout transforms. The drop size of voxels are set to 30% of input sub-volumes. We randomly generate masked ROIs inside image, until the total masked voxels are larger than scheduled number of dropping voxels. Unlike canonical pre-training rules of masked tokens in BERT [85], our local transformations to the CT sub-volumes are then arranged to neighbouring tokens. This scheme can construct semantic targets across partitioned tokens, which is critical in medical spatial context. By analogy to Models Genesis [2], which is CNN-based model consisting expensive convolutional, transposed convolution layers and skip connection between encoder and decoder, our pre-training approach is trained to reconstruct input sub-volumes from the output tokens of the Swin Transformer. Overall, the intuition of modeling inpainting, rotation prediction and contrastive coding is to generalize better representations from aspects of images context, geometry and similarity, respectively.

CHAPTER 9

Characterizing Renal Structures with 3D Block Aggregate Transformers

9.1 Introduction

Hierarchical models [42, 115, 127] are received significant interest in medical image analysis due to their advantages of modeling heterogeneous high-resolution radiography images. Recent works on vision transformers [88, 204] show superior performance on visual representations compared to state-of-the-art convolution-based networks [223]. However, ViT usually requires large-scale training data with expensive clinical expertise [127, 224]. When trained on smaller cohorts, transformer-based models often suffer from a lack of inductive bias [88, 254] and lead to data inefficiency. Moreover, the self-attention mechanism on modeling multi-scale features for high-resolution medical volumes is computationally expensive [204, 255, 256]. These challenges inspire designing hierarchical transformer structures in analogy to convolution-based networks (e.g., 3D UNet [115]). Addressing the data inefficiency of transformers is critical for its application on medical image analysis, especially for understanding small targets on 3D high-dimensional image volumes.

One such challenge is segmenting the small structures of kidney sub-components. Renal structure volumes from clinical CT scans have been recently suggested as a useful surrogate for evaluating renal function [257, 258]. These investigations elucidate the correlations of the volumetric measurements on the renal cortex, medulla, and pelvicalyceal system with kidney function. In such studies, manual segmentation is performed as the gold standard for visual and quantitative morphological assessment on CT scans [259] as shown in Figure 9.1. However, manual quantification by clinical experts is resource-intensive, time-consuming, and may suffer from insufficient inter- and intra-reproducibility.

To improve the representation learning of transformers in small datasets, recent works envision the use of local self-attention to form hierarchical transformers [204, 230, 256]. To leverage information across embedded sequences, "shifted window" transformers [204] are proposed for dense predictions and modeling multi-scale features. However, these attempts that aim to complicate the self-attention range often yield high computation complexity and data inefficiency. Inspired by the

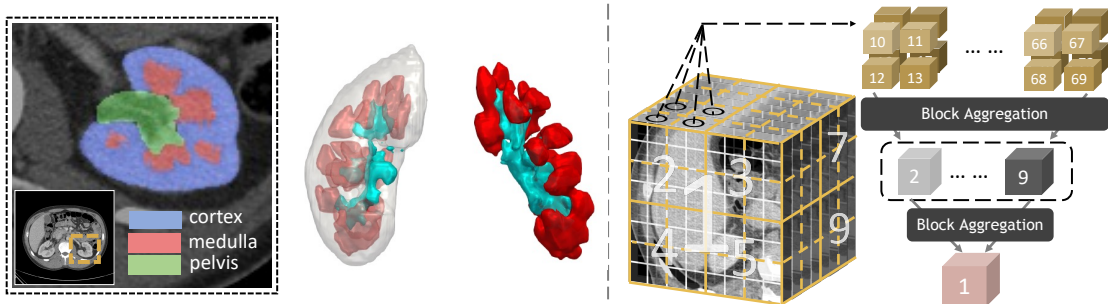


Figure 9.1: Left: visual and 3D illustration of the kidney components. Right: Demonstration of the hierarchical transformer design, the 3D block aggregation is conducted every two hierarchies, blocks at a factor of 8 are merged to perform communication of sequence representations.

aggregation function in the nested ViT [260], we propose a new design of a 3D U-shape medical segmentation model with Nested Transformers (UNesT) hierarchically with the 3D block aggregation function, that learn locality behaviors for small structures or small dataset. This design retains the original global self-attention mechanism and achieves information communication across patches by stacking transformer encoders hierarchically.

Our contributions in this work can be summarized as:

- We introduce a novel 3D medical segmentation model, named UNesT with a 3D block aggregation function. This method achieves hierarchical modeling of high-resolution medical images and outperforms local self-attention variants with a simplified design, which leads to improved data efficiency.
- We collect and manually delineate the first renal substructures dataset (116 patients) on characterizing multiple kidney components. We show that our method achieves state-of-the-art performance to accurately measure the cortical, medullary, and pelvicalyceal system volumes.
- We demonstrate the clinical utility of this work by accurate volumetric analysis, strong correlation, and reproducibility. Validation on external public dataset KiTS shows the generalizability of the proposed method.

9.2 Related Works

3D Medical Segmentation with Transformers. Transformer-based 3D medical image segmentation models [93, 214, 224, 261, 262, 263, 264, 265, 266] are popular and achieve state-of-the-

art performance in several benchmarks. The self-attention mechanism [128] allows the inputs at different positions of a sequence to interact with each other, and then compute the overall representation from the sequences. Although transformers exhibit outstanding performance in learning global context, their deficiency in capturing localized information remain. To address this, TransFuse [262], TransBTS [261], CoTr [93], UNETR [214] are proposed architectures which combine transformers and CNNs into hybrid designs. More recently, hierarchical transformers are proposed with shifted-window [204], it enables cross-patch self-attention connections. Based on Swin ViT, Swin UNETR [127, 267] and SwinUNET [230] are introduced for capturing multi-scale features in CT images. However, the modification on local self-attention results in quadratic increase of complexity.

Hierarchical Feature Aggregation. The aggregation of multi-level features could improve the segmentation results by merging the features extracted from different layers. Modeling hierarchical features, such as U-Net [115] and pyramid networks [42], multi-scale representations are leveraged. The extended feature pyramids compound the spatial and semantic information through two structures, iterative deep layer aggregation which fuses multi-scale information as well as hierarchical deep aggregation which fuses representations across channels. In addition to single network, nested UNets [268], nnUNets [58], coarse-to-fine [185] and Random Patch [43] suggest multi-stage pathways that enrich the different semantic levels of feature progressively with cascaded networks. Different from the above CNN-based methods, we explore the use of data-efficient transformers for modeling hierarchical 3D features by the block aggregation.

9.3 Method

9.3.1 UNesT Architecture

The proposed network contains a hierarchical transformer as the encoder, which consists of three hierarchies to perform self-attention communications among image blocks. Following the motivation of NesT [260] for natural images, we process the volumetric information between 3D adjacent blocks by the aggregation layer every two hierarchies. The overall architecture, as shown in Figure 9.2, also contains skip connections with convolution modules and a decoder for better capturing localized information.

Given the input image sub-volume $\mathcal{X} \in \mathbb{R}^{H \times W \times D}$, the volumetric embedding token is with

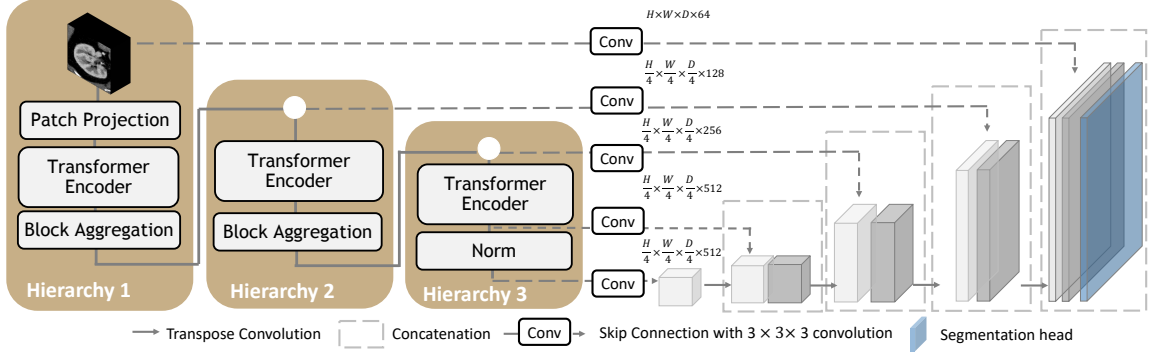


Figure 9.2: Overview of the proposed UNesT with the hierarchical transformer encoder. Block aggregation and image feature down-sampling are performed between hierarchies.

patch size of $S_h \times S_w \times S_d$. Then all projected sequences of embeddings are partitioned to blocks with a resolution of $\mathcal{X} \in \mathbb{R}^{b \times T \times n}$, where T is the number of blocks at the current hierarchy, b is the batch size, n is the total length of sequences. The dimensions of the embeddings follow $T \times n = \frac{H}{S_h} \times \frac{W}{S_w} \times \frac{D}{S_d}$. In the subsequent transformer layers, we use the canonical multi-head self-attention (MSA), multi-layer perceptron (MLP), and Layer normalization (LN). We add learnable position embeddings to sequences for capturing spatial relations before the blocked transformers. The output of encoder layers $t - 1$ and t are computed as follows:

$$\begin{aligned} \hat{z}^t &= \text{MSA}_{\text{HRCHY}_l}(\text{LN}(z^{t-1})) + z^{t-1} \\ z^t &= \text{MLP}(\text{LN}(\hat{z}^t)) + \hat{z}^t, \end{aligned} \quad (9.1)$$

where $\text{MSA}_{\text{HRCHY}_l}$ denotes the multi-head self-attention layer of hierarchy l , \hat{z}^t and z^t are the output representations of MSA and MLP. In the practice, $\text{MSA}_{\text{HRCHY}_l}$ is applied parallel to all partitioned blocks:

$$\begin{aligned} \text{MSA}_{\text{HRCHY}_l}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{Stack}(\text{BLK}_1, \dots, \text{BLK}_T) \\ \text{BLK} &= \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{\sigma}}\right)\mathbf{V}, \end{aligned} \quad (9.2)$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ denote queries, keys, and values vectors in the multi-head attention, σ is the size of each vector. All blocks at each level of hierarchy share the same parameters given the input \mathcal{X} , which leads to hierarchical representations without increasing complexity. Finally, the block aggregation is merged spatially by adjacent 8 blocks.

9.3.2 3D Block Aggregation

Following [260], we extend the spatial nesting operations to 3D blocks where each volume block is modeled independently. Information across blocks is communicated by the aggregation module. At hierarchy l , the spatial operations are conducted to down-sampled feature maps at $\mathbb{R}^{b \times H'/2 \times W'/2 \times D'/2}$. At the bottom of each hierarchy, the embeddings are blocked back to feature $Z_{l+1} \in \mathbb{R}^{b \times T/8 \times n}$ for hierarchy $l + 1$. There are three hierarchies in our model design, a factor of 8 is reduced in a total number of blocks which results in [64, 8, 1] blocks. In the volumetric plane, the encoded blocks are merged among adjacent blocks representations. The design and use of the aggregation modules in the 3D scenario leverage local attention, lead to a data-efficient design.

9.3.3 Decoder

To better capture localized information and further reduce the effect of lacking inductive bias of transformer, we use a hybrid design with a convolution-based decoder for segmentation. The features from different hierarchies of the transformer encoder are fed into skip connections followed by convolution layers. As shown in Figure 9.2, we extract the output representations at the image level and each hierarchy to $3 \times 3 \times 3$ conv layers, then upsample by a factor of 2. Next, the output of the transposed conv is concatenated with the prior hierarchy representations. The segmentation mask is acquired by $1 \times 1 \times 1$ conv layer with a softmax activation function. Compared to some prior related works such as TransBTS [261] and CoTr [93], our design employs the hierarchical transformer directly on images and extract representations at multiple scales without conv layers.

9.4 Experiments

9.4.1 Dataset

Renal Substructure Dataset. The study design uses clinically collected renal CT of 116 de-identified patients accessed under IRB approval. We use selected ICD codes related to kidney dysfunction as exclusion criteria, that could have a potential influence on kidney anatomies. The left and right renal structures are outlined manually by three interpreters under the supervision of clinical experts. The annotation for the cortex label also includes the renal columns, the medulla is surrounded by the cortex, and the pelvicalyceal systems contain calyces and pelvis that drain into the ureter. All manual labels are verified and corrected independently by expert observers. For the

Table 9.1: Segmentation results of the renal substructure on testing cases. The UNesT achieves state-of-the-art performance compared to prior kidney components studies and 3D medical segmentation baselines. The number of parameters and GFLOPS (with a single input volume of $96 \times 96 \times 96$) are shown for deep learning-based approaches. * indicates statistically significant ($p < 0.01$) by Wilcoxon signed-rank test.

Method	#Param	GFLOPS	Cortex		Medulla		Pelvicalyceal System		Avg.	
			DSC	HD	DSC	HD	DSC	HD	DSC	HD
Chen et al. [270]	N/A	N/A	0.7512	40.1947	N/A	N/A	N/A	N/A	N/A	N/A
Xiang et al. [271]	N/A	N/A	0.8196	27.1455	N/A	N/A	N/A	N/A	N/A	N/A
Jin et al. [272]	N/A	N/A	0.8041	34.5170	0.7186	32.1059	0.6473	39.9125	0.7233	35.5118
Tang et al. [123]	40.9M	423.9	0.8601	19.7508	0.7884	18.6030	0.7490	34.1723	0.7991	24.1754
nnUNet [58]	19.1M (3DUNet)	412.7	0.8915	17.3764	0.8002	18.3132	0.7309	31.3501	0.8075	22.3466
TransBTS [261]	33.0M	359.4	0.8901	17.0213	0.8013	17.3084	0.7305	30.8745	0.8073	21.7347
CoTr [93]	46.5M	399.2	0.8958	16.4904	0.8019	16.5934	0.7393	30.1282	0.8123	21.0707
nnFormer [224]	158.9M	146.5	0.9094	15.5839	0.8104	15.9412	0.7418	29.4407	0.8205	20.3219
UNETR [214]	92.6M	41.2	0.9072	15.9829	0.8221	14.9555	0.7632	27.4703	0.8308	19.4696
UNesT	87.3M	37.5	0.9201	14.5401	0.8356	13.5933	0.7843	24.5445	0.8467*	17.5593

test set of 20 subjects, we perform a second round of manual segmentation (interpreter 2) to assess the intra-rater variability and reproducibility.

KiTS19. To validate the generalizability of the proposed method while remaining the target of characterizing renal tissues, we apply the model to the public KiTS19 dataset. The KiTS19 [269] task focuses on the whole kidney and tumor segmentation. We perform five-fold cross-validation experiments and show results of the held-out 20% as testing.

9.4.2 Implementation Details

Five-fold cross-validation is used for all experiments on 96 subjects, while 20 subjects are used for held-out testing. The five-fold models’ ensemble is used for inferencing and evaluating test set performance. For experiment training, we used 1) CT window range of $[-175, 275]$ HU; 2) scaled intensities of $[0.0, 1.0]$; 3) training with single Nvidia RTX 2080 11GB GPU with Pytorch and MONAI implementation at batch size of 1 (input image sub-volume size of $96 \times 96 \times 96$); 4) AdamW optimizer with warm-up cosine scheduler of 500 steps. The learning rate is initialized to 0.001 followed by a decay of $1e^{-5}$ for 50K iterations. For fair comparison and direct evaluation of the effectiveness of models, no pre-training is performed for all segmentation tasks.

Metrics. Segmentation performance is evaluated between ground truth (rater 1) and prediction by Dice-Sorensen coefficient (DSC), and symmetric Hausdorff Distance (HD). Volumetric analyses are evaluated under R squared error, Pearson R, absolute deviation of volume, and the percentage difference between the proposed method and manual label.

Table 9.2: Comparison of volumetric analysis metrics between the proposed method and the state-of-the-art clinical study on kidney components.

Metrics	Cortex		Medulla		Pelvicalyceal System	
	Tang et al. [123]	UNesT	Tang et al. [123]	UNesT	Tang et al. [123]	UNesT
R Squared	0.9200	0.9359	0.6652	0.6837	0.4586	0.5917
Pearson R	0.9838	0.9891	0.8156	0.8368	0.6772	0.7148
Absolute Deviation of Volume	3.0233	2.7254	3.5496	3.2958	0.9443	0.8012
Percentage Difference	4.8280	3.9478	7.4750	7.0382	19.0716	13.5737

9.5 Results

9.5.1 Characterization of Renal Structures

We evaluate the UNesT performance on two groups of methods: 1) the clinical kidney components studies such as CortexSeg [270], CorteXpert [271], AAM [272], and 2) recent conv- [58] and transformer-based [93, 214, 224, 261] 3D medical segmentation baselines.

Segmentation Results. Compared to canonical kidney studies using shape model or random forests in Table 9.1, the deep learning-based methods improve the performance by a large margin from 0.7233 to 0.7991. Among the nnUNet [58] and extensive transformer models, we obtain the state-of-the-art average Dice score of 0.8467 compared to the second-best performance of 0.8308, with a significant improvement $p < 0.01$ under Wilcoxon signed-rank test. We observe higher improvement on smaller anatomies such as medulla and collecting systems. We compare qualitative results in Figure 9.3. Our method demonstrates the distinct improvement of detailed structures for medulla and pelvicalyceal systems.

Volumetric Analysis. Table 9.2 lists the volume measurement with the proposed method. The UNesT achieves an R squared error of 0.9359 on the cortex. The correlation performance metric with Pearson R achieves 0.9891 for the UNesT against the manual label on the cortex. Our method obtains 2.7254 with an absolute deviation of volumes. The percent difference in the cortex is 3.9478. Quantitative results show that our workflow can serve as the state-of-the-art volumetric measurement compared to prior kidney characterization pipeline [123].

9.5.2 Ablation Study

Effect of the Block Aggregation. We show the hierarchical architecture design (with 3D block aggregation) is critical for medical image segmentation (as shown in Figure 9.4 left and middle). The result shows that the hierarchy mechanism achieves superior performance at 20% to 100% of

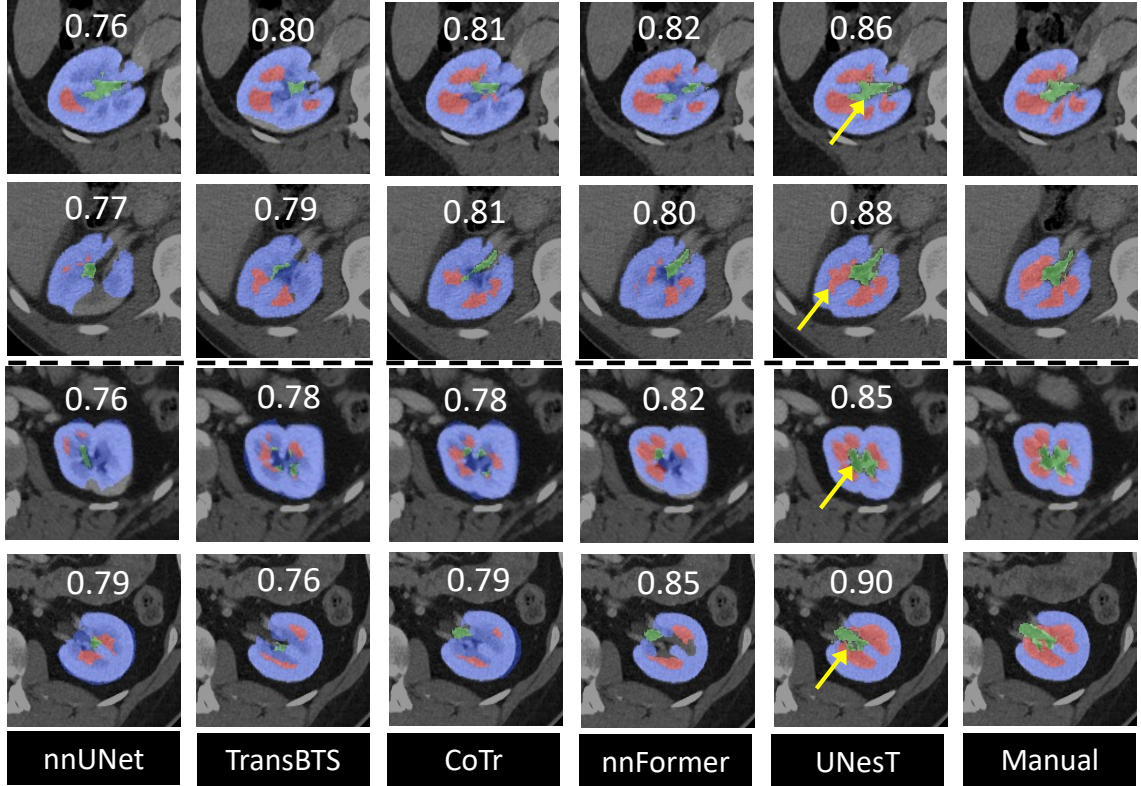


Figure 9.3: Qualitative comparisons of representative renal sub-structures segmentation on two right (top) and two left (bottom) kidneys. The average DSC is marked on each image. UNesT shows distinct improvement on the medulla (red) and pelvicalyceal system (green) against baselines.

training data. At the low data regime, the block aggregation achieves a higher improvement ($> 4\%$ of DSC) compared to the second-best method. We notice that the model without block aggregation (canonical transformer layers) obtains lower performance. The results show that block aggregation performs as a critical component for representation learning for transformer-based models.

Data Efficiency. The Figure 9.4 shows the data efficiency of our proposed method. First, UNesT achieves better performance when training with fewer data. Second, UNesT with block aggregation demonstrates a faster convergence rate (15% and 4% difference at 2K/30K iterations) compared to the backbone model without hierarchies.

Generalizability. To validate the generalizability of the UNesT, we compare KiTS19 results among nnUNet [58] and transformer-based methods. Our approach achieves moderate improvement at DSC of 0.9778 and 0.8398 for kidneys and tumors, indicating that the designed architecture can be used as a generic 3D segmentation method.

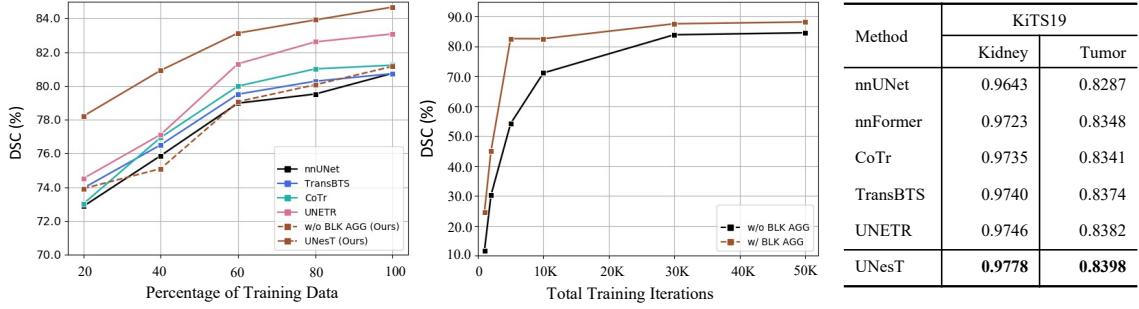


Figure 9.4: Left: DSC comparison on the test set at different percentages of training samples. Middle: Comparison of the convergence rate for the proposed method with and without hierarchical modules, validation DSC along training iterations are demonstrated. Right: Results on the KiTS19 dataset show the generalizability of the proposed UNesT.

9.6 Discussion and Conclusion

In this paper, we target the critical problem that transformer-based models are commonly data-inefficient, which leads to unsatisfied performance when tasked with learning small structures and small datasets. In this work, we develop the first cohort of renal sub-structures study, specifically the renal cortex, medulla, and pelvicalyceal system. Upon the clinically acquired subjects, we propose a novel hierarchical transformer-based 3D medical image segmentation approach (UNesT). We show that the proposed method is data-efficient for accurately quantifying kidney components and can be used for volumetric analysis such as the medullary pyramids. Figure 9.5 in the supplementary materials shows the proposed automatic segmentation method achieves better agreement compared to inter-rater assessment, 0.01 against 0.29 of mean difference indicating reliable reproducibility.

Clinical Impact. Visual quantitative analysis of renal structures remains a complex task for radiologists. Some of the histomorphometry features of regions of the kidney (e.g. textural or graph features) are poorly adapted for manual identifications. In this study, we show that UNesT achieves consistently reliable performance. Compared with previous studies on cortex segmentation, the proposed approach significantly facilitates derivation of the visual and quantitative results.

Efficient segmentation is critical for clinical practice in deploying individual assessment. We note that, unlike other large organs, the renal segmentation dataset can be different in terms of imaging protocols, patient morphology, and institutional variations. We consider the framework adaptable to the segmentation of abnormal primitives in the future. In terms of sensitivity, we believe that the approach can be further improved from two perspectives. First, pre-registration of the

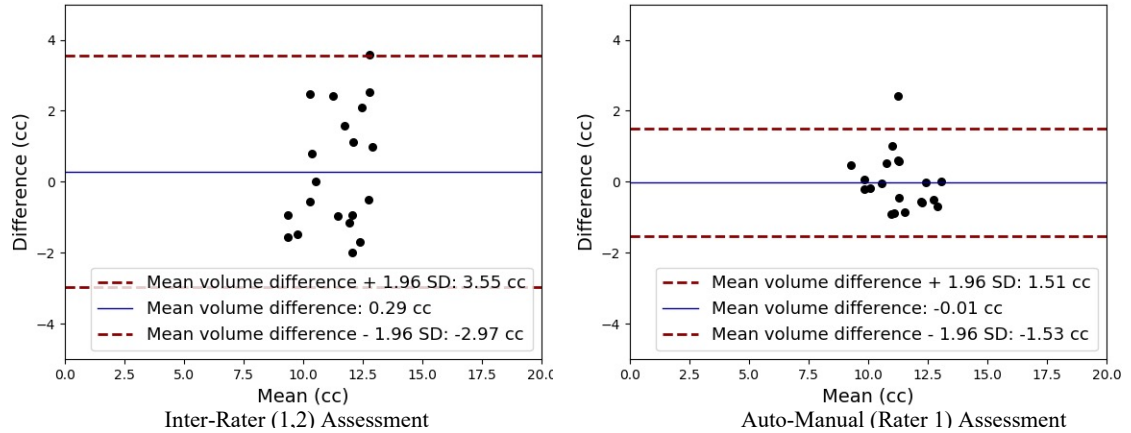


Figure 9.5: The Bland-Atman plots compare the medulla volume agreement of inter-rater and auto-manual assessment. We show the cross-validation on interpreter 1, interpreter 2 manual segmentation on the same test set. Interpreters present independent observation without communication. The auto-manual assessment shows the agreement between UNesT and interpreter 1 annotations.

kidney region of interest can help to reduce the shape and size variations and thus boost the segmentation performances. Second, incorporating dose usage in the segmentation loop can be very helpful. It can be expected that augmented contrast can be measured to better identify adjacent tissues among renal structures.

CHAPTER 10

Abdominal Atlas Template for Accurate Tissue Correspondence

10.1 Introduction

The construction of human body templates can provide information of anatomical details of organs and neighboring tissues. The study of body atlas can differentiate and contextualize findings of target organs. However, currently there is no standard reference space of abdomen to define platforms for healthy anatomies. Complicated relationships between cells in multi-perspectives such as organization, specialization and cooperation, are challenging to analyze. Extensive studies of mapping the organization and molecular profiles of cells within tissues or organs are needed across the human body. While the majority of efforts are distributed to the cellular and molecular perspectives [273], generalizing information from cell to organ level is essential to provide a better understanding of the functionality and linkage across scales. The use of computed tomography provides an opportunity to contextualize the anatomical characteristics of organs and systems in the human body. Creating a generalizable framework with integration of micro-scale information and system-scale information benefits clinicians and researchers to visualize details across scales. The linkage between the correspondence of the organ anatomical structures to the pathogenesis of organ-related disease can be explored, increasing the confidence level of the explainable findings with the use of the anatomical information generated from the generalizable framework [274].

In this work, we present a contrast-preserving CT abdominal atlas framework, optimized for healthy kidney organs with contrast-characterized variability and the generalizable features across a large population of clinical cohorts. Specifically, we initially extracted the abdominal volume of interest with a similar field of view to the atlas target image, using a deep neural network called body part regression (BPR) [125]. 2D slices of the CT volume assessed with BPR model and generate a value ranging from -12 to +12, corresponding to an approximate anatomical location in the body. By limiting the range of values for abdominal regions, each CT volume is cropped and excludes other regions apart from the abdomen, such as the lung and pelvis. A two-stage hierarchical registration pipeline is then performed, registering the extracted volume interest to the high-resolution atlas

target with affine and non-rigid registration [275, 276]. To ensure the stability and the variation localized in the atlas template, average and variance mappings across the multi-contrast registered output is computed to demonstrate a better understanding of anatomical details of kidney organs across different contrasts.

Besides the whole kidney, Renal structure volumes from clinical CT images have been recently suggested as a useful surrogate for evaluating renal function [277, 278]. These investigations established correlations of measurements of the renal cortex, medulla, and pelvicalyceal system with kidney function. Manual segmentation remains the gold standard for visual and quantitative morphological assessment on arterial phase CT scans [279, 280]. However, manual quantification by kidney experts is resource-intensive, time consuming, and may suffer from insufficient inter- and intra-reproducibility.

The increasing demands for useful large-scale volume quantification presents opportunities and challenges to develop rapid and accurate methods for measuring renal structures. Imaging hardware, acquisition protocols, and non-uniform shapes can lead to significant variability in renal morphology. Though imaging modalities such as computed tomography (CT) and magnetic resonance imaging (MRI) provide high quality images, the renal cortex, medulla and pelvicalyceal system can be difficult to differentiate due to low-contrast boundaries. As shown in Figure. 10.1, contrast-enhanced CT can provide increased imaging contrast and can characterize some aspects of renal physiology. The arterial phase CT scans show the optimal contrast between the enhanced cortex, medulla, and pelvicalyceal system. Prior investigations focused on renal cortex segmentation by CT and MRI [278, 280]. However, these studies were limited to a small number of datasets and required tedious manual assessment. Consequently, semi-automatic studies on images done for clinical indications were established for exploring risk factors and markers in the setting of kidney disease [281, 282, 283, 284]. Because potential renal structure delineation undergo a standardized evaluation in the contrast-enhanced CT, fully-automatic evaluation methods were feasible and useful for characterizing clinical correlations in the larger population [282, 283, 285, 286]. Among these, Chen et al. [285] proposed a method with active appearance model (AAM) for cortex segmentation on abdominal CT. Jin et al. [282] extensively studied AAM that combines random forest and 3D Hough transform. Xiang et al. [283] constructed a cortex model that adapted model with graph search. However, the cortex, medulla, and pelvicalyceal system contain different nephron

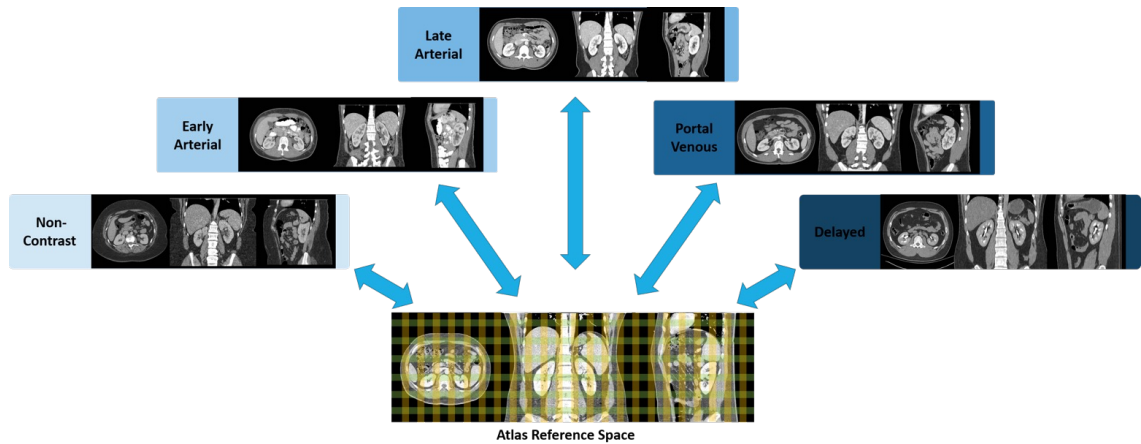


Figure 10.1: Illustration of defining atlas standardize reference for multi-contrast phase CT. The color grid in the three-dimensional atlas space represents the defined spatial reference for the abdominal volume of interest and localize abdominal organs with each contrast phase characteristics. Blue arrows represent the bi-directional transformation across the atlas target defined spatial reference and the original source image space.

segments that may have different markers in assessment of renal function. Separate analysis of cortex, medulla and the collecting system may help clarify associations with clinical factors and enable discrepant findings. To date, study of simultaneously segmenting renal cortex, medulla and pelvicalyceal system in contrast-enhanced abdominal CT has only rarely been explored.

In addition, we also show the construction of pancreas atlas, the first healthy pancreas atlas template for tissue mappings.

10.2 Data

Kidney and Pancreas Atlas: A large clinical cohort of multi-contrast CT was employed for abdominal organ registration. In total, 2000 patients' de-identified CT data were initially retrieved in de-identified form from ImageVU with the approval of Institutional Review Board (IRB). For kidney atlas, in these 2000 patients, since some of them consisted of kidney organ disease scans, qualified criteria in ICD-9 codes and age range from 18-50 years old were set and applied to extract scans with healthy kidneys from all subjects. 720 subjects out of 2000 were identified after quality assessment and extract the corresponding contrast phase abdominal CT scans, which included 290 unlabeled CT volumes in total with: 1) non-contrast: 50 volumes, 2) early arterial: 30, 3) late arterial: 80 volumes, 4) portal venous: 100 volumes, 5) delayed: 30 volumes. For pancreas atlas, a same

query process is performed, but to keep normal pancreas subjects for building the atlas template. All CT volumes are initially reoriented to standard orientation before further processing [178]. BPRN was initially performed to each modality volumes and obtain the similar field of view with the atlas target. They were then resampled to the same resolution and dimensions with the atlas target for performing registration pipeline. We aim to adapt a generalized atlas framework for localizing the anatomical and contextual characteristics of kidney organs across multi-contrast.

Renal Structures: The patient dataset (102 subjects) used in our experiments was acquired from Vanderbilt University Medical Center (VUMC) archival clinically obtained images with IRB approval. The initial query for this study included 2000 patients. As shown in Figure. 10.2, to obtain comprehensive evaluation, we assessed each subject's deidentified electronic health record (EHR) data, including demographics, ICD-9, and ICD-10 codes. The aim of the assessment was to identify normal kidneys from the large-scale cohort. We used selected ICD codes related to kidney dysfunction as exclusion criteria, supervised by kidney experts, that could have potential influence for kidney anatomies. A small subset of subjects was categorized by ICD-10 codes; these ICD-10 codes were converted to ICD-9 standard. For this control study, we selected adult subjects 18 to 50 years old, (median age of 35). Of 2000 potential subjects, 720 subjects were retrieved after exclusion for ICD codes indicating possible kidney dysfunction. We further limited our study to those who had arterial abdominal CT scans, which yielded 102 subjects. As depicted in Figure. 10.2, we split the 102 subjects into several subsets for different usage. The left and right renal structures were outlined manually by interpreters under supervision of clinical experts. On CT scans in the dataset 1, testing set (dataset 3), the renal cortex, medulla and pelvicalyceal systems were segmented by the first interpreter using a tool for academic usage, a tracing pad is used for delineating contours. For each tissue, the interpreter was instructed to verify the segmentation slice-by-slice in all axial, sagittal, and coronal views. The annotation for cortex label also includes the renal columns, the medulla was surrounded by the cortex, and the pelvicalyceal systems contain calyces and pelvis that drain into ureter. All manual labels were verified and corrected independently by expert observers. Dataset 1 was used as training and evaluation for the base automatic system. Afterward, the dataset 2 labels are initialized by applying the first system, and then manually refined to reach the standard of manual segmentation (on average 1 hour per subject for refinement). Last, we trained the final automatic system that incorporates all subjects. For the test set of 20 subjects (dataset 3), we performed a

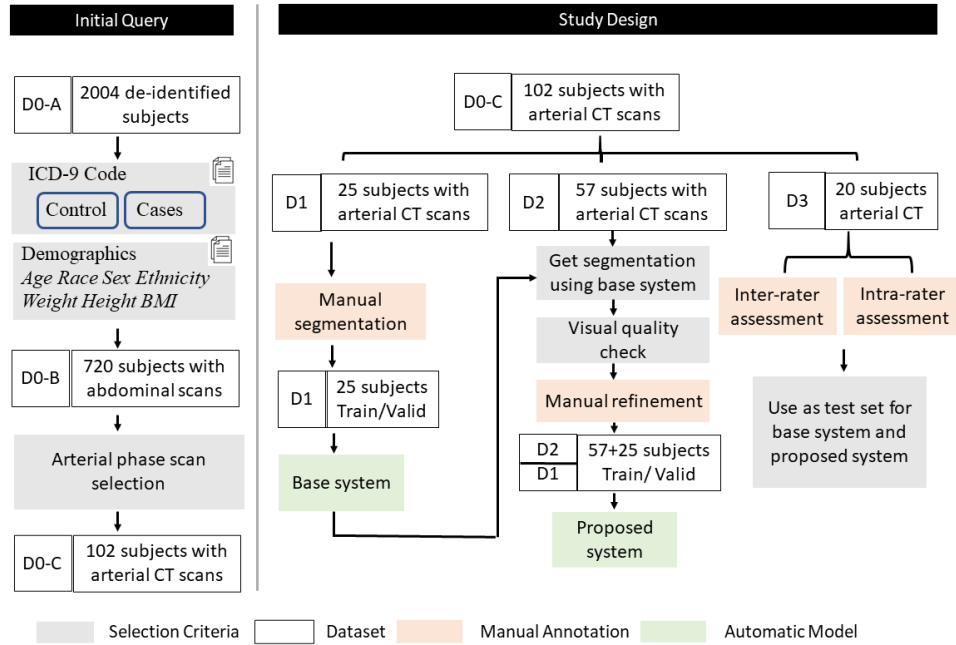


Figure 10.2: Framework of study design. From D0-A to D0-C, the flowchart shows the criteria from initial query to the CT scans used in this study. Dataset 1 was used for training and validating base system, and dataset 3 is held for external testing. Dataset 1 and dataset 2 were used for training and validation the proposed system.

second round of manual segmentation to assess the intra-rater variability. The same procedure was used to label three renal structures. To assess inter-rater variability and perform reproducibility, we have another two interpreters perform manual segmentations on a subset in dataset 3 independently using the same annotation protocol, as shown in Figure. 10.2. We evaluated the intra- and inter-rater variability on this dataset by Bland-Altman agreement plots in the result section. Both interpreters are supervised by experienced radiologists (> 10 years of experience in kidney radiology).

10.3 Experiments

10.3.1 Kidney and Pancreas Atlas

On constructing the kidney and pancreas atlas, we performed registrations, each source image (moving/floating) is resampled to the same voxel resolution and the 3-dimensional shape of the atlas target (fixed/target) with cropping/padding slices. Our registration pipeline is composed of 2 hierarchal stages: 1) affine registration and 2) non-rigid registration. Dense displacement sampling (DEEDS), a 3D medical registration tool in discretized sampling space that has been shown to yield a great

performance in abdominal organs registration, is used for both affine non-rigid registration in this pipeline. The DEEDS affine registration is first performed to initially align both moving images and the atlas target to preserve 12 degrees of freedom of transformation and provide a prior definition of the spatial context and each affine component. An affine transformation matrix is generated as the output and become the second stage non-rigid registrations' input. The DEEDS non-rigid registration is refined with the spatial context as the local voxel-wise correspondence with its specific similarity metric, which will be illustrated below. Five different scale levels are used with grid spacing ranging from 8 to 4 voxels to extract patches and displacement search radii from six to two steps between 5 and 1 voxels. Deformed scans with the displacement data from randomly control point selection is generated and localize the original image space voxel information to the atlas space after deformation. To ensure the stability of the atlas generated, all successfully deformed scans are summed and compute an average and variance mapping to visualize the intensity fluctuation and variation around the abdominal body and kidney/pancreas organs.

The similarity metric defined in DEEDS registration tool is measuring the self-similarity context with the patches extracted from moving images. Such a similarity metric aims to find a similar context around neighboring voxels in patches. The self-similarity metric S is optimizing a distance function D between the image patches extracted from the moving image M . A function q^2 is computed to estimate both the noise in local and global perspectives. Meanwhile, a certain number of neighborhoods kinds N is also defined as to determine the kinds of self-similarities in the neighborhood. As randomly extracting an image patch with a center at x , the measurement calculation of the self-similarity can be demonstrated as follows:

$$D(M, x, y) = \exp\left(\frac{S(x, y)}{q^2}\right) \quad (10.1)$$

where y is defined as the center of another patch from one of the neighborhood N . This similarity metric provides an opportunity to avoid the image artifacts or random noise from the central patch extracted and prevent a direct adverse effect in calculation. The distance between pairwise patches is calculated within six neighborhoods and concentrate in extracting the contextual neighboring information, instead of the direct shape representation.

10.3.2 Automated Renal Structures Segmentation

The automatic segmentation is based on deep neural networks. We trained the system in two stages in order to capture both global and local context, and to prevent instability due to resolution changes, as shown in Figure. 10.3. In the first stage, we used the entire volume of each CT scans, each down-sampled to (2 2 6) mm and padding to volume size of 168 168 64. The purpose of the first phase is to capture the global context and locate the relative position of kidneys/pancreas. We used the 3D U-Net [115] as the backbone, a state-of-the-art segmentation network. While there are other backbone architectures, these are expected comparable empirical performance. In the second stage, we randomly select predicted voxels in the first stage. Using the pixel indices as the center, a renal bounding box was calculated and created. In order to perform randomness, a random shift was added. The distance of shifting was derived by a random number generator with mean and variance among bounding boxes centers, as shown in Figure. 10.3. Then, we trained a patch-based network using the same backbone network, followed by majority voting. In our experiment, we used Dice loss as the error function to adapt the multi-labels in the training. We also used instance normalization due to the batch size of 1 employed in the experiment, each normalization layer is followed by a Rectified Linear Unit (ReLU).

Experimental Design. Original CT scans were clipped by soft tissue window with a range from -175 to 250 Hounsfield Unit (HU). Second, the soft-tissue windowed CT images and labels were resampled and normalized using cubic spline and nearest interpolation, respectively. All abdominal scans were processed using body part regression technique to ensure the same field of view (FOV) from middle lung to pelvis. We performed 5-fold cross validation during experiments, 82 subjects (dataset 1 and dataset 2) are uniformly split to five-folds, for each experiment, 4 folds were used for training and the remaining one set for validation. This procedure was repeated five times until every set was used for validation. The stopping criteria and optimal model were selected according to the best performance in validation set. The best performed model on validation set was selected for testing. The withheld 20 subjects (dataset 3) were used as the testing set.

Statistical Analyses. Renal cortex, medulla and pelvicalyceal systems volumes were computed as the product of the number of foreground pixels multiplied with the pixel spacing in x and y direction and the corresponding slice thickness. Accuracy. Several metrics have been proposed to evaluate imaging segmentation accuracy. We computed the Dice coefficient and average Hausdorff

Table 10.1: Summarized statistics for the final automatic system compared to manual segmentation.

	Cortex	Medulla	Pelvicalyceal System
Dice similarity coefficient	0.8701±0.0382	0.7984±0.0425	0.7590±0.04486
Hausdorff Distance	19.7508±10.3980	18.6030±9.0224	34.1723±8.0668
R Squared	0.9200	0.6652	0.4586
Pearson R	0.9838	0.8156	0.6772
Absolute deviation of volume (cm ³)	3.0233±2.2435	3.5496±1.8402	0.9443±1.0004
Percent difference (%)	4.8280±3.6849	7.4750±6.9467	19.0716±21.0213

Table 10.2: Comparison between our best performance and state-of-the-art segmentation methods.

Method	Dice Score			Hausdorff Distance			Absolute Deviation of Volume			reference
	cortex	medulla	PS	cortex	medulla	PS	cortex	medulla	PS	
Chen et al.	0.7512	None	None	40.1947	None	None	12.6296	None	None	[172]
Xiang et al.	0.8196	None	None	27.1455	None	None	8.8125	None	None	[170]
Jin et al.	0.8041	0.7186	0.6473	34.5170	32.1059	39.9125	9.2454	8.8578	3.5491	[169]
Ours	0.8701	0.7984	0.7590	19.7508	18.6030	34.1723	3.0233	3.5496	0.9443	

PS: pelvicalyceal system

distance to judge segmentation from different methods against the ground truth. We compared the different methods' volume predictions against the ground truth with Dice scores, Hausdorff distance, and absolute deviation of volumes.

Agreement and Bias. Bland-Altman plots serve to compare different algorithm's estimates agreement with the manual segmentation.

10.4 Results

We show the qualitative representations as shown in Figure. 10.3 of registrations across multiple contrast phases are shown in terms of single subject registration, average template and variance mapping in Figure. 10.3 and the surface tissue mapping (Figure. 10.4). The multi-organ label provided general localization information of anatomical structures in atlas target image and compared with the deformed output from each contrast phase. The qualitative representation of the single subject registration demonstrated good localization ability of kidneys/pancreas organ with the corresponding contrast characteristics, while the kidney/pancreas organs in the early and late arterial phase subject is shown with small level of over-deformed. The average mapping of each contrast phase was then computed with all successfully registered contrast-corresponding volumes. Each

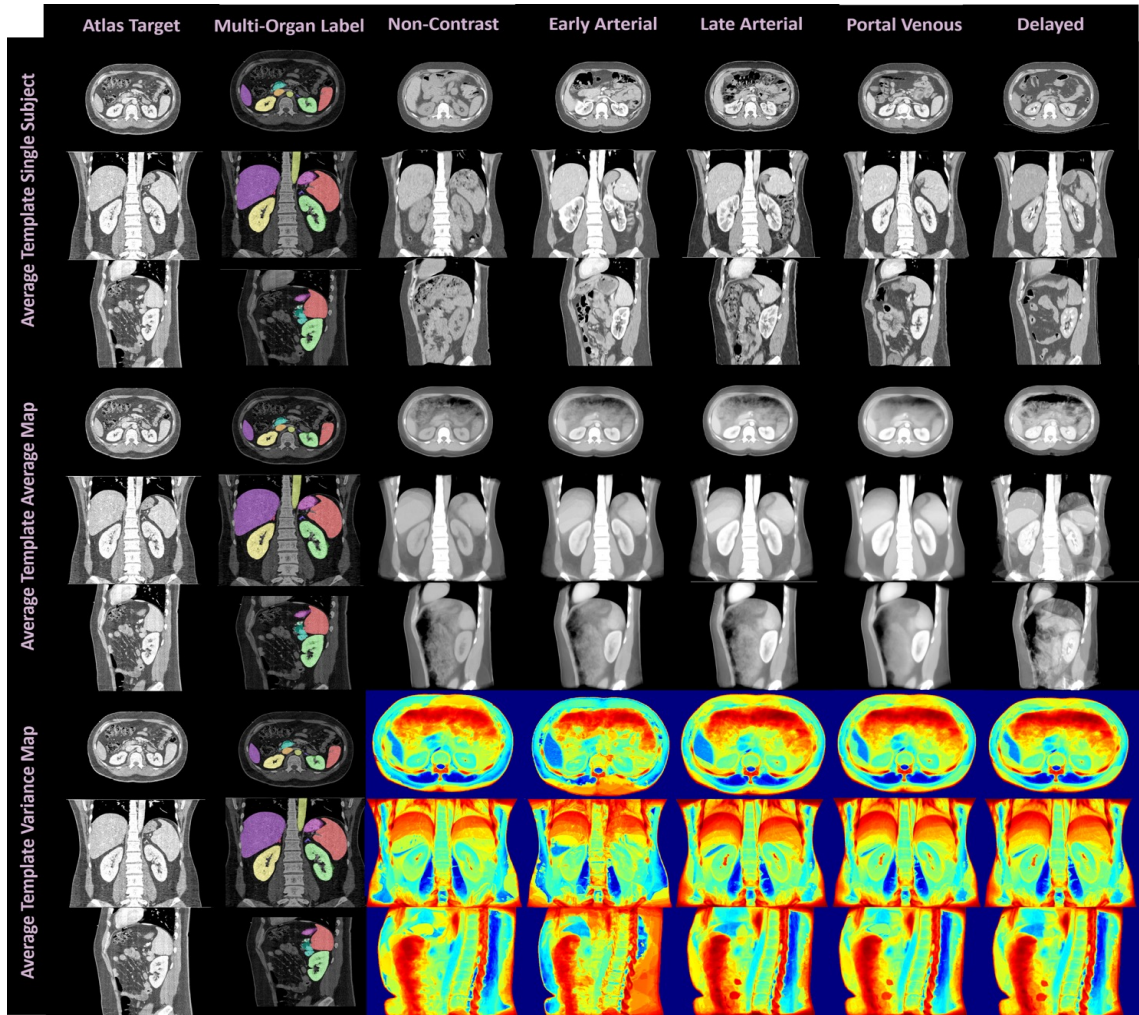


Figure 10.3: The qualitative representation of the single subject registrations, average mapping and variance mapping of each contrast phases are demonstrated. The contrastive and morphological characteristics of kidney organs are demonstrated in the single subject registration and average mapping of each phases. Small variations are shown surrounding the kidney organs region in the variance mapping, while great variations are located in the diaphragm region nearby with liver and spleen (Color bar is provided in the supplementary file).

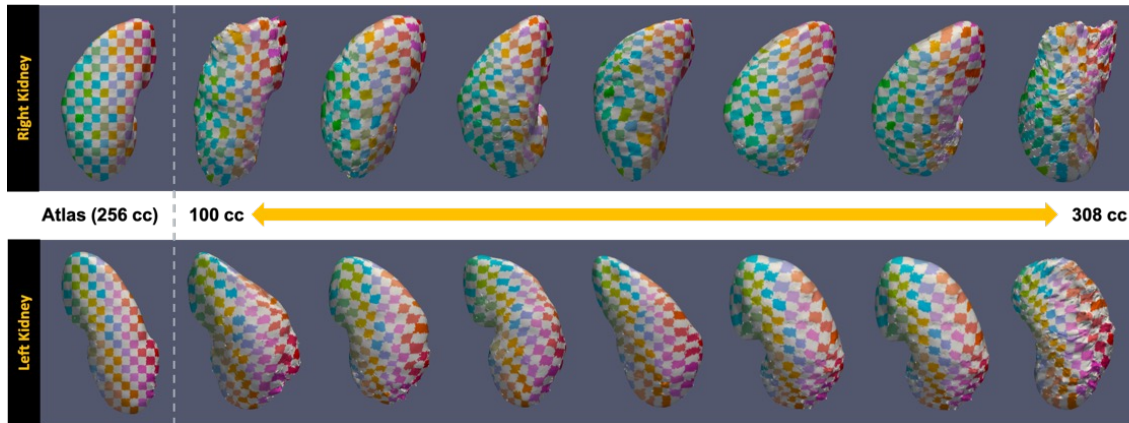


Figure 10.4: The surface rendering of the registered kidney with significant morphological variation are also illustrated. The 2D checkerboard pattern demonstrate the correspondence of the deformation from atlas space to the moving image space. A stable deformation across the change in volumetric morphology of kidney (100 cc to 308 cc) are demonstrated with the deformed checkerboard.

average template is shown with stable abdominal body registration and a significant clear boundary of both left and right kidneys. The contrast characteristics of the kidneys were demonstrated and well allocated in the similar anatomical location of the atlas template. Apart from the contrast characteristics, the anatomy of kidney sub-structure and renal-related vessels can barely appear in the average template of early arterial, late arterial and delayed phase. To ensure the stability of transferring kidneys' anatomical information, variance map of each contrast phase template is also computed to demonstrate the voxel variability of each organ across the clinical cohort. The small variation in the kidney is illustrated with a color range from yellow to green, while significant variation in voxels is shown near the diaphragm region and the color range from orange to red indicated the highly deformed variability across the registered outputs. Overall, the anatomical and contrast characteristics of kidney organs can be both preserved and transferred to the atlas template with good stability.

In Figure. 10.6, we show the constructed pancreas average template and its segmentation atlas fused by queried population data. The qualitative results show the stability of the proposed framework for pancreas atlas. In Figure. 10.7, the surface rendering of the 3D checkerboard show the consistent correspondence of the pancreas across different subjects.

Renal Segmentation Similarity Analysis. We compared our automatic renal segmentation methods on the “leave-out” test set (dataset 3 in Figure. 10.2). The mean Dice and variance are

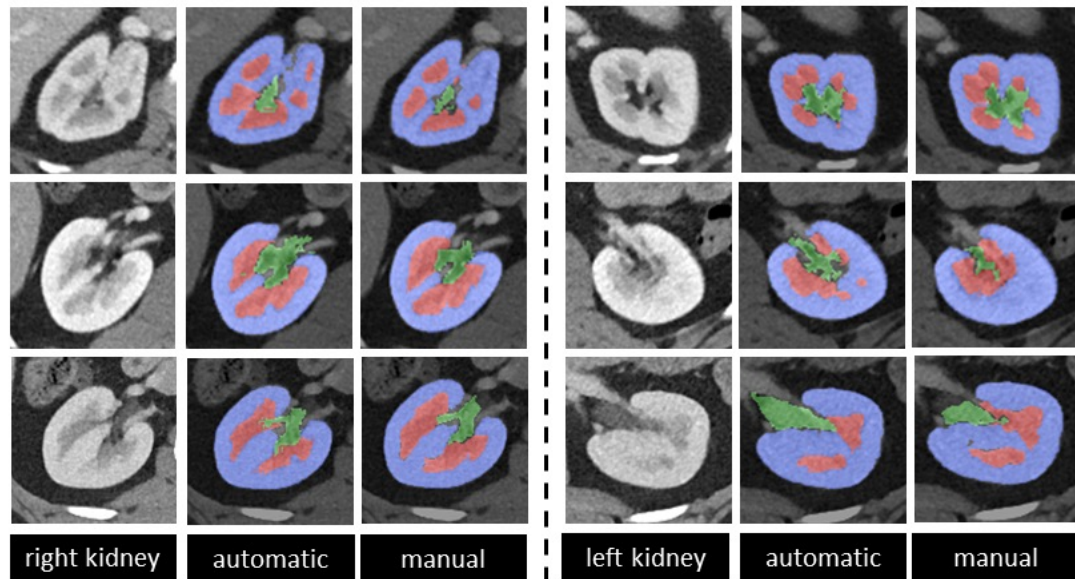


Figure 10.5: Examples of representative subjects' visualization. Sections surrounded by blue labels are cortex segmentation, the red is the medulla label, and green is the collecting system. The automatic segmentation are acquired from using the final system model, the comparison between automatic result with manual annotations shows that our method achieves comparable segmentation performance that can be used for measurements.

0.8701 (cortex), 0.7984 (medulla), and 0.7590 (pelvicalyceal system). We calculated the Hausdorff Distance (HD) for comparing the segmentation result, the HD metric evaluates the largest surface distance between automatic and manual labels. The pelvicalyceal system shown received on HD of 34.1723. Dice scores of 0.7512 and 0.8196 (p-value < 0.001, statistically significant in. The relative absolute volume difference showed that our method achieves 3.0233. On the basis of metrics, our method outperformed previous studies as shown in Table. 10.2. Representative segmentation visualization of the automatic model is shown in Figure. 10.5.

We calculated the volume measurement by the automatic method. The automatic method achieves R squared error of 0.9200 on cortex. The correlation performance metric with Pearson R achieved 0.9838 for the automatic system against manual label on cortex. The automatic method obtained 3.00233 with absolute deviation of volumes. The percent difference on cortex volumes is 4.8280. Metrics are evaluated for medulla and pelvicalyceal system. Quantitative results showed that the automatic model could be used as the automatic volume measurement for the qualitative observer method.

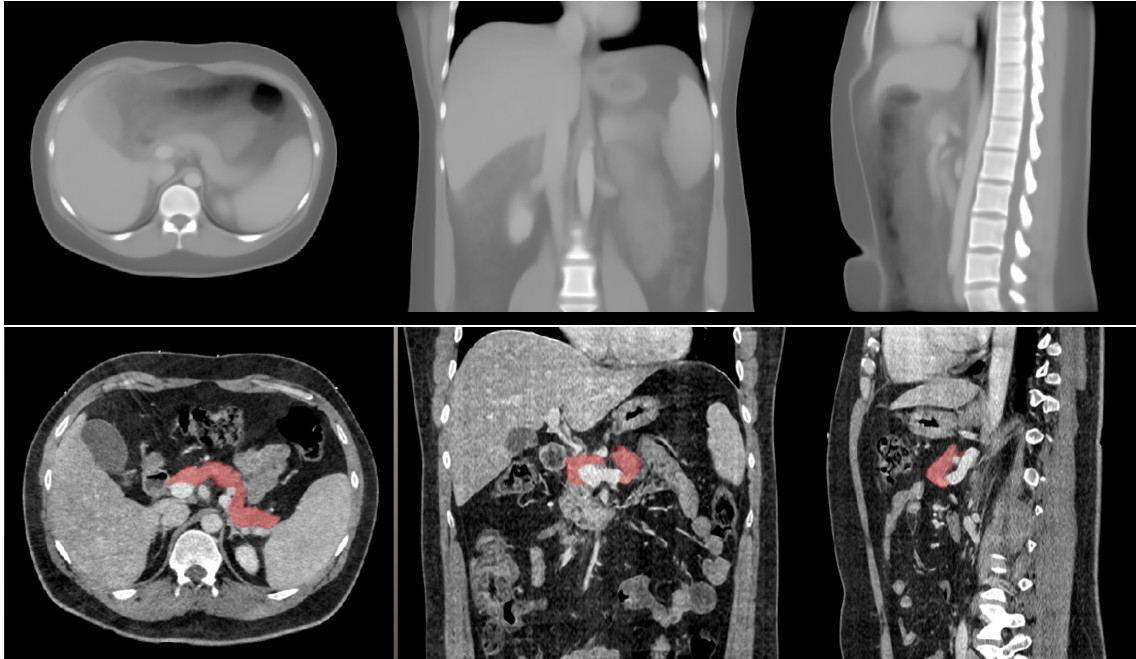


Figure 10.6: Average template for pancreas atlas. The segmentation atlas in red color is shown in three planner view.



Figure 10.7: The surface rendering of the registered pancreas with variations. The 3D checkerboard pattern demonstrates the correspondence of the deformation from atlas space to the moving image space.

10.5 Conclusion

This section presents a healthy kidney atlas to generalize the contrastive and morphological characteristics across patients with significant variability in demographics and imaging protocols. Specifically, the healthy kidney atlas provides a stable reference standard for both left and right kidney organs in 3-dimensional space to transfer kidney information using registration pipeline. Significant variance on the field of view and the organ shape can be focused as the optimization parameters to reduce the possibility of failure registration. Potential future exploration with the use of the atlas template can be further investigated in both engineering and clinical perspectives, to provide better understandings and measures towards the kidney organs.

Additionally, we developed an automatic algorithm to segment the renal cortex, medulla, and pelvicalyceal system using deep neural networks in arterial phase CT scan. From the clinical acquired dataset, this study shows that this automatic method to segment renal structures achieves reliable performance. We also show that the automatic method is feasible in quantifying imaging of potential biomarkers in complex scenarios such as the medullary pyramids.

Visual quantitative analysis of renal structures remains a complex task for radiologists. Some of the histomorphometry features of regions of the kidney (e.g. textural or graph features) are sub-visual for human eyes. In this study, we showed that our proposed framework achieved consistently high performance. Compared with previous studies on cortex segmentation, the deep neural network segmentation approach significantly facilitated derivation of the visual and quantitative result.

In the volumetric analysis, the R squared error and Pearson R results in Table. 10.1 show strong correlation between automatic volume measurements and manual references, especially in the renal cortex and medulla. In this study, we established the approach of assessing the repeatability and reproducibility of the manual effort across three independent observers, and we achieved consistent agreements. This effort contributed to assess the automated segmentation algorithm that can yield the same result for a given image set after training with these ground truth data. We observed a relatively small volume difference between the performance of automatic system and manual ground truth for renal cortex and medulla, while the pelvicalyceal system showed a higher difference, reflecting this very challenging tissue to assess from images.

In line with previous literature focused on cortex, such as studies that measured volumes for assessing kidney donation and studies on renal segmentation for MR Urography, we achieved seg-

mentation on cortex, medulla, and pelvicalyceal system simultaneously in arterial CT scans. Our study suggests that fully automatic medulla segmentation is feasible and accurate. In evaluating time efficiency, manual segmentation took about 7 hours per CT scan. The acceleration may provide a rationale and further support for delivering treatment in clinical procedures.

CHAPTER 11

Clinical Application: Validation and Estimation of Spleen Volume Via Computer-assisted Segmentation on Clinically Acquired CT Scans

11.1 Introduction

Organ size measurements provide clinical utility in diagnosis and assessing treatment response with different cancers. For instance, reduction in spleen volume is a crucial response element for myeloproliferative neoplasms (MPN) [3], and recent studies showed that significant (greater than or equal to 35%) splenic volume reduction (SVR) in myelofibrosis patients is associated with improved overall survival [287, 288]. Splenic volume is measured in MPNs as an indicator of extramedullary hematopoiesis, and SVR is a surrogate endpoint used to measure success of an intervention in MPNs [289, 290]. Quantitative estimates of splenic biomarkers have been of clinical interest in molecular, histologic, radiographic, and physiologic characterization of spleens [15, 16, 291]. Unique challenges emerge when validating machine learning generated imaging biomarkers. Changes in imaging hardware, acquisition parameters, and patient positioning during imaging cause significant variations among images. Several groups, such as the NCI-sponsored Quantitative Imaging Network [292] and the Radiological Society of North America's Quantitative Imaging Biomarker Alliance [293], have pursued multicenter imaging trials to address these variabilities. Using machine learning algorithms to compute biomarkers presents additional challenges. Performance from machine learning algorithms is dependent not only on the training data set but also data set upon which the algorithms are deployed. Park and Han addressed this challenge in their recent evaluation of diagnostic and predictive artificial intelligence technologies. They postulate that validating such algorithms' performance requires testing in a clinical cohort that adequately represents the target patient population [294]. Given an image from ultrasound, magnetic resonance imaging or computed tomography (CT), manual annotation of spleen segmentation is still the gold standard [5], yet the process is time intensive and requires domain expertise. Here, computer-assisted spleen labeling could not only reduce resource consumption, but also support investigation into spleen size's clinical utility as a biomarker [19, 154, 295]. Several challenges remain, including

significant inter-subject variability in spleen size, shape, and orientation. The multi-atlas approaches produced encouraging results [296, 297], and we recently developed a deep convolutional neural network algorithm that performs significantly better than previous methods [141]. The non-manual measurements are “fit for purpose”; i.e., the rigor and methods utilized should be in accordance with the intended purpose of the biomarker study [298]. Given this “fit for purpose” approach, it is difficult to establish absolute benchmark criteria for validation studies. Accordingly, the reporting of validation studies becomes increasingly important to allow investigators to evaluate whether the assay is appropriate for a proposed study. We hypothesize that automated measures of spleen size can serve as a biomarker for clinicians to better predict MPN disease progression and response to therapy. Investigation into and potential utility of spleen volume as a biomarker is limited due to the time-intensive process in obtaining those measurements. Indeed, Sargent et al. defined a set of prerequisite criteria for an imaging biomarker before clinical validation studies can be performed [299]. Notably, the authors highlight the necessity that the technology to assess the biomarker of interest be stable and widely available. With automated spleen volume estimation, it is essential to validate the methods following such criteria for clinical validation. The goal of this study is to evaluate the performance of the proposed state-of-the-art spleen segmentation algorithm in measuring spleen volumes on clinically acquired CT scans from patients with MPNs. Our validation study includes a complete assessment of the technical performance of the biomarker assay using the state-of-the-art deep neural networks. Such assessment includes measures of assay accuracy, repeatability, reproducibility, technical bias, sensitivity, and specificity. We investigate four pipeline estimates for using the deep learning algorithms on all scans: 1) manual segmentation by expert readers; 2) automatic segmentation using deep learning algorithms; 3) unidimensional measurements and 4) 3-D splenic index measurement. Further, the validation study defines the limits of the detection and quantification for a given assay [15]. In the context of imaging biomarkers, the assay includes both the image generation process (the specific imaging protocol) and the subsequent post-processing procedures to yield the biomarker measurement accuracy. Ultimately, through the validation process, we show the agreement and bias proportion to the intended purpose of the biomarker study. In the cross-validation experiments, our computer-assisted method produced segmentation masks with an averaged dice coefficient of 0.95148 when evaluating against hand labeled masks. Most importantly, our proposed method’s volume estimation achieved R2 coefficient of 0.99800 and Pearson R coeffi-

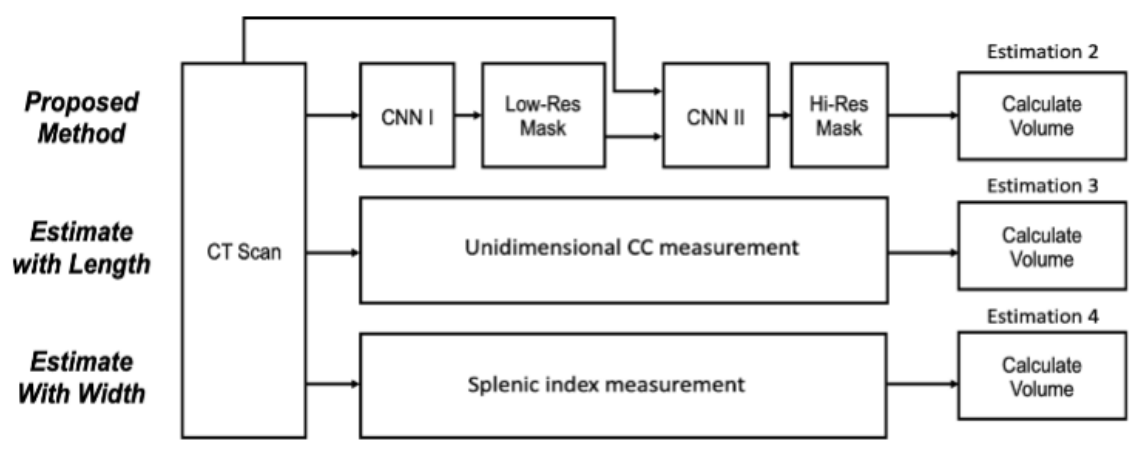


Figure 11.1: Pipeline for the proposed method and unidimensional linear regression estimation methods. The computer-assisted method in estimation 2 includes two CNN models with a coarse-to-fine framework. Estimation 3 and 4 use measurements of length and width (splenic index) from the ground truth for cc (cubic centimeter) volume estimation.

cient of 0.99905, indicating a significant improvement over traditional linear estimations [45]. This demonstrates the potential of obtaining more accurate spleen volume estimates from state-of-the-art deep learning algorithm.

11.2 Materials and Methods

Data Acquisition Under Institutional Review Board (IRB) approval, we obtained 138 de-identified abdominal CT from patients enrolled in NCT02493530. This is a Phase 1 multi-center study of TGR-1202 administered together with ruxolitinib in patients with MPNs. To minimize spectrum bias, we utilized a consecutive series of patients from four study locations (Mayo, Wisconsin, Colorado, Vanderbilt). Including multi-center data provides evidence to evaluate how this algorithm adapts to the variability inherent in multi-center trials. The spleen was segmented by expert readers from each scan in this dataset to establish the ground truth.

Manual Segmentation (Estimation 1)

Manual spleen segmentation on all 138 scans establishes baseline splenic volumes. We used open sourced tool MIPAV software from the NIH (21) to trace the spleen anatomies. In our study, CT scans from patients with splenomegaly were retrieved. We delineated the outlines on every axial slice and filled the regions enclosed by the tool. A radiologist, certified by the abdominal imaging board, verified all splenic contours on the volumetric investigations. We calculated ground truth

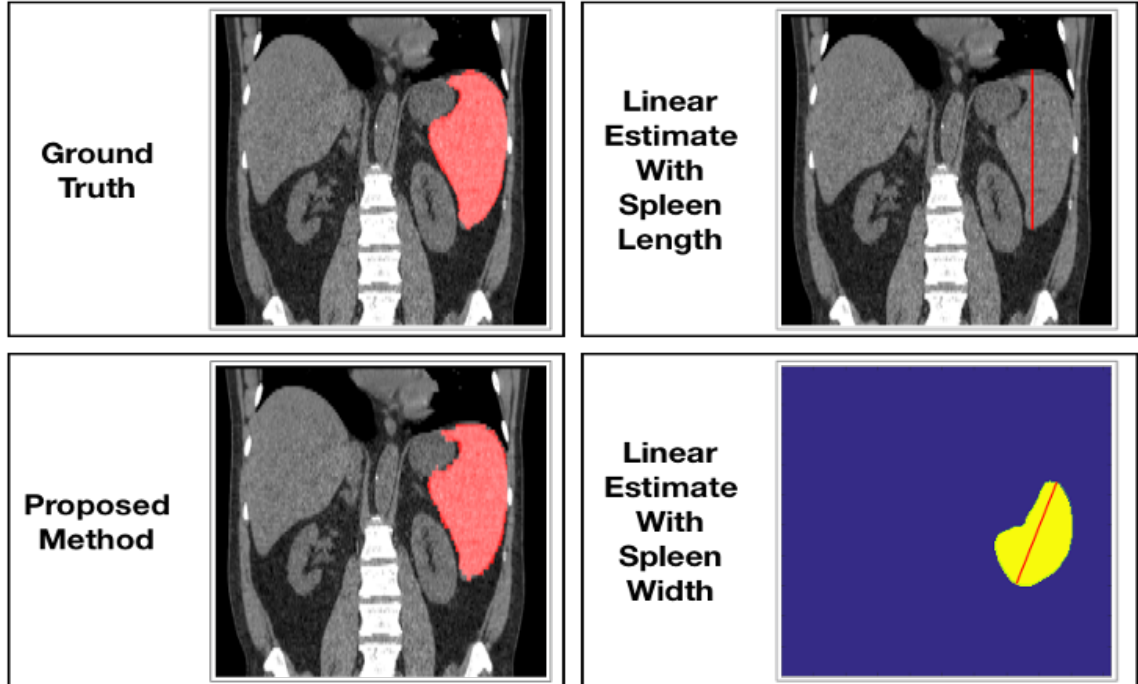


Figure 11.2: Demonstration of the measurements from pipelines for estimating spleen volumes. The manual and computer-assisted methods evaluate the spleen volume (estimation 1 and 2). The linear estimates (3 and 4) manually extract splenic diameters along different axes (length and width) from an unlabeled CT scan.

spleen volume for each image by directly multiplying unit volume (cc/voxel) with number of voxels inside segmentation region.

In order to evaluate the repeatability and reproducibility, we retrieved a subset of 40 patients labeled by a second similarly qualified imaging analyst under the supervision of a radiologist to assess the inter-rater reliability of manual segmentation. Both readers adhered to the same tracing protocol, and the agreement evaluation is shown in the result by the Bland-Atman plot (Figure. 11.5). **Deep Convolutional Neural Network Algorithm (Estimation 2)**

Stage1: Low-resolution segmentation. Given CT scans with a fine resolution of $[0.8 \times 0.8 \times 2\text{mm}]$, we first downsampled the images and trained a 3D U-Net for segmentation with lower resolution. Each scan slice is downsampled from $[512 \times 512]$ to $[168, 168]$ and images were normalized to a consistent voxel resolution of $[2 \times 2 \times 6\text{mm}]$. We used Dice loss to compare network outputs and ground truth labels. We ignored the background loss in order to increase weights for anatomies. The crude segmentation masks are then upsampled to original resolution with nearest interpolation for later stages. This approach is trained end-to-end, and Figure. 11.3 to Figure. 11.5 assess of the

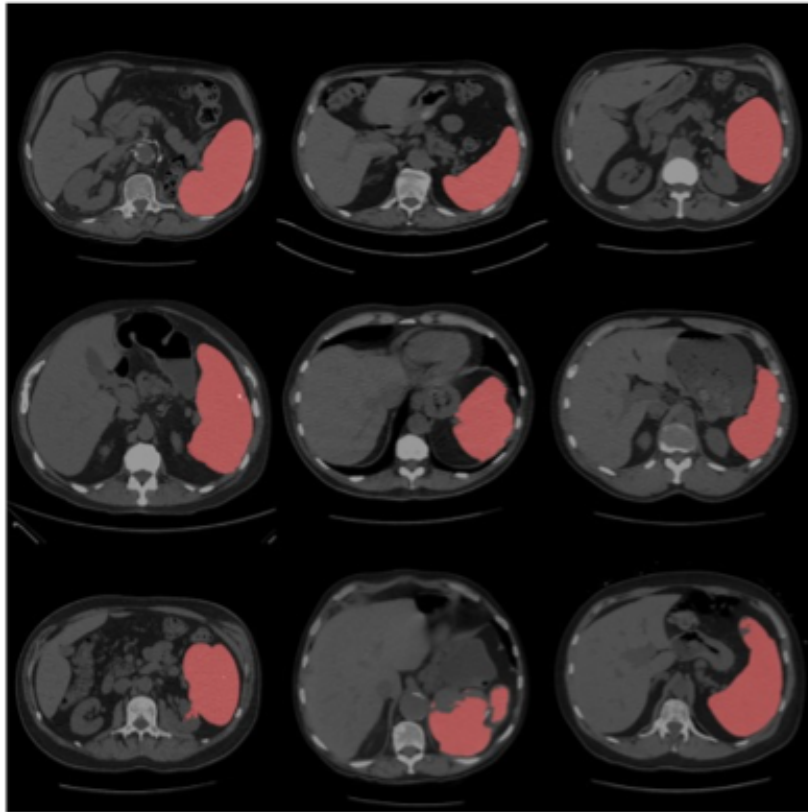


Figure 11.3: Quality assurance of the deep learning method in estimation 2 with computed tomography. Top row: three representative subjects' slice above state-of-the-art. Middle row: three representative cases with successful segmentation. Bottom row: failure cases where manual correction was required.

approach’s segmentation quality. The down-sampled volume in low-resolution framework while lacking detailed structures of anatomies, still preserves complete spatial context in CT scan.

Stage2: Random patch selection. For each CT scan, we randomly selected voxels in the predicted coarse segmentation mask. Fixing the selected voxels as centers, we placed bounding cubes with slight random shifts along all axes. A Gaussian random variable determines the shifting distance. High resolution patches from original images were cropped according to bounding cubes, and they formed second stage model inputs (Middle panel of Figure.11.1). This strategy builds the hierarchy of non-linear features from random patches regardless of 3D contexts, and it employs detailed context at original resolution and incorporates advantages of data augmentation with shifting.

Stage3: High-resolution segmentation and label fusion. Using randomly selected high-resolution patches from the prior stage, we trained a second 3D U-Net. Integrating all patches on field of views, we estimated the full field of view. Majority vote is used to merge estimates into a final segmentation yielding the spleen voxels. Specifically, after separating full spatial context to randomly selected subspaces, the overlapped regions provide more than one segmentation label for a voxel. We summarize a single label given a vector of class labels from candidates. We ignore voters outside the image space and related values are excluded in the label fusion.

Linear Estimation of Spleen Volume with Spleen Length (Estimation 3 and 4)

Recently, Bezerra et al. [45] introduced a helpful approach that estimates spleen volume through unidimensional spleen measurements and 3-D splenic index. We obtained maximum spleen length from coronal/frontal plane (L) and maximum spleen width from oblique sagittal/axial plane (W). We calculated spleen volume estimates with linear regression equations specified by Bezerra et al. [45]. The equations they proposed for maximum spleen length and maximum spleen widths are $V = (L - 5.8006)/0.0126$ and $V = (W - 8.1101)/0.0098$. Figure. 11.2 depicts maximum spleen length and maximum spleen width alongside other estimation methods in this study.

11.3 Analysis.

Accuracy

Several metrics have been proposed to evaluate imaging segmentation accuracy [300]. We computed the Dice coefficient and average Hausdorff distance to judge segmentation from different methods against the ground truth (Table. 11.1). We compared the different methods’ volume pre-

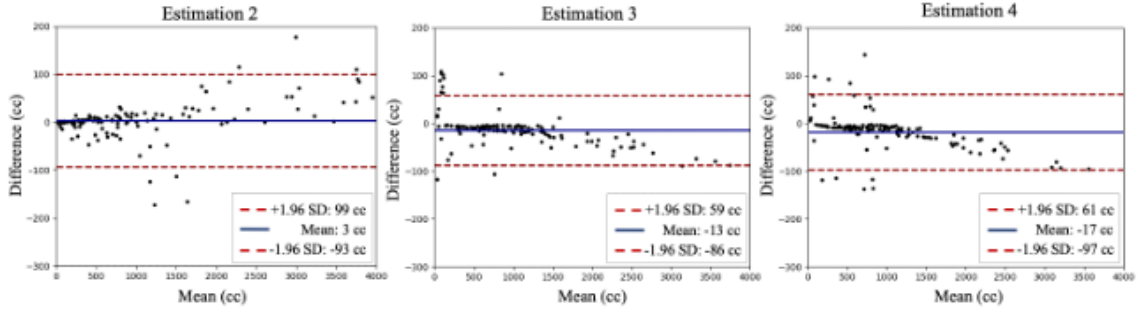


Figure 11.4: Bland-Altman Plot for computer-assisted method (estimation 2), linear estimate with length and splenic index (estimation 3 and 4). On each plot, the x-axis indicates the mean volume between the ground truth and the estimation from computer-aided method. The y-axis shows the difference in volume. 1.96 standard deviation is shown as the confidence interval.

dictions against the ground truth with R-squared, Pearson correlation, and absolute and percent deviations.

Bias

Bland-Altman plots (Figure. 11.4) serve to compare different algorithm’s estimates’ agreement with the ground truth. (Figure. 11.4).

Reproducibility

Once trained, the automated segmentation algorithm yields same result for a given image. To ascertain the method as a sound replacement for manual segmentation, we examined manual the approach’s reproducibility on a subset of 40 patients. This subset of patient images was labeled simultaneously by a second research associate in order to assess the inter-rater reliability. Assuming the label from expert 2 is the ground truth, we present the reproducibility comparisons in Figure. 11.5.

11.4 Results

In Figure 11.3, we present examples of predicted segmentation masks on their respective CT scans. The top row shows three examples with exceptional alignment. The second row’s predictions are satisfactory but slightly flawed at the edges, and the bottom row presents some of our failure cases.

As shown in Table. 11.1, our proposed method’s population Dice coefficient statistic is 0.9515 ± 0.0332 , indicating a high degree of alignment between prediction masks and ground truth masks. The averaged symmetric Hausdorff Distance is 9.3846 ± 15.113 . The superior volume estimation performance of the proposed method becomes more apparent when we compare estimation 2’s R2

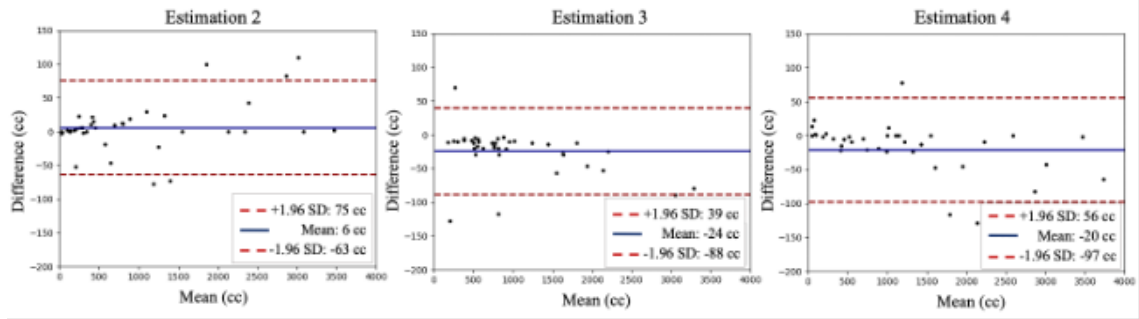


Figure 11.5: The repeatability and reproducibility between different imaging analyst readers on 40 respective studies. Bland-Altman Plot between estimations (2-4). The mean in difference and a confidence interval of 1.96 standard deviation are shown.

Table 11.1: Summarized Statistics for different estimations compared to ground truth

	Proposed Method (Estimation 2)	Linear Estimation with Length (Estimation 3)	Linear Estimation with Width (Estimation 4)
Dice similarity coefficient	0.951±0.033	N/A	N/A
Hausdorff Distance	9.385±15.113	N/A	N/A
R Squared	0.998	0.954	0.973
Pearson R	0.999	0.963	0.978
Absolute deviation of volume (cm ³)	16.718±24.504	20.481±38.195	26.815±40.576
Percent difference (%)	1.892±2.975	4.125±4.981	4.015±5.573

values (0.99800), Pearson R coefficient (0.99905), absolute difference in volume estimates (29.091 ± 113.720 cm³) and percent difference in volume estimates (2.3443 ± 6.2031 cm³) (p < 0.05 with paired t-test between estimation 2 and 3, 2 and 4 respectively) against those of the unidimensional linear regression estimation methods. The Bland-Altman plots in Figure. 11.4 also illustrates that the deep convolutional neural network algorithm produced superior results.

11.5 Discussion

Main Contributions The study demonstrates a deep learning algorithm’s ability to produce precise spleen masks and spleen volume estimates from abdominal CT scans. Shown in Bland-Altman plots, estimates 2-4 performing on normal patient CT scans and failed to inference accurately where splenomegaly was present. Our proposed method yielded superior results in both cases and was also able to produce empty mask for CT scans after splenectomy. Shown in Table. 11.1, our method achieves a Pearson R coefficient of 0.99905 and an average non-significant absolute deviation of

20.604cc with respect to the ground truth. This approach performs consistent comparable result with resource-intensive manual segmentation with unidimensional measurements. These results show that a deep learning algorithm supervised by manual segmentation can enable generation of higher accuracy in estimation of spleen volumes.

In reviewing the performance of inter-rater reproducibility, we found that agreement between experts were highly accurate (mean 6 cc in Bland-Atman plot in Figure. 11.5). Using the labeling by the second expert as the ground truth, the computer-assisted segmentation achieves slightly higher agreement (mean 3cc in Figure. 11.4). Typically, the automatic method observes outliers include spleens with severe splenomegaly and those under surgery, which these outliers are required by rudimentary visual quality refinement.

11.6 Clinical Improvement

Organ size measurements remain attractive biomarkers for the assessment of disease. However, diagnosis time is always a concern, and the significant time and resource cost associated with the extraction of organ size limits its use. So far, such methods demonstrated limited clinical utility to justify its adoption in clinical workflows, and it is our vision that automated measures of organ size will reduce the cost of obtaining such measures, allowing for the prospective evaluation of organ sizes in the study of disease prognosis and treatment response.

11.7 Summary

In summary, we proposed a deep convolutional neural network algorithm that produced more accurate spleen volume estimates for abdominal CT scans. Given the importance of spleen volume as a biomarker and considering the superior effectiveness of the algorithm on patients with splenomegaly, we conclude the algorithm can provide sufficiently accurate spleen volume measurements. Given the current inaccurate and computationally expensive algorithms or accurate but laborious manual labeling, this proposed method should exempt expert radiologists from arduous manual labeling of spleens while allowing more precise, and expedient clinical diagnosis and treatment suggestions with better spleen volume estimates.

CHAPTER 12

Conclusion and Future Works

12.1 Impact of the Dissertation

This dissertation applies the data-driven perspectives of statistical, machine learning, and deep learning models in healthcare. Starting from fundamental medical image analysis techniques such as data curation, segmentation, registration, synthesis and classification, we investigate several most challenging problems for 3D clinical-acquired images, especially in abdomen imaging.

We had close collaborations with Vanderbilt University Medical Center (VUMC), by querying from ImageVU, we extract thousands of subjects with abdomen CT images ($> 10,000$ CT volumes). The first big challenge is to identify research quality data from the large-scale clinical-acquired cohorts. To automatically identify the data within region of interest, we propose the body part regression technique (**BUSN (Chapter 2)** and **SemiBR (Chapter 3)**). The tool can efficiently detect slice-wise localization. Moreover, we show the curation tool can effectively clean data, which used for down-streaming tasks such as segmentation. With the self- and semi-supervised learning technique, we incorporate the deep learning with large-scale data of multi-contrast, multi-institutes, and multi-regions. The other challenging task for accessing CT images is to identify enhancement phases, the contrast phase for CT images is critical as it highlight the target tissues for screening. Also, there is distribution shifting if the machine learning model is biased to single or several contrast phase data. To address the problem, we propose the GAN-based method (**CDGAN (Chapter 4)**) to identify the enhancement phases of raw clinical-acquired images. In summary, we target to create the automatic image quality assurance tools for CT images, and applies to large-scale data-driven machine learning techniques to effectively extract research quality data.

Quantifying medical images and measuring biomarker metrics can be useful for delivering clinical decisions and research discoveries. One of the most crucial mean for quantitative measurement is the segmentation. Medical image segmentation, especially for 3D CT and MRI volumes, are challenging because the high-resolution, high-dimensional, and complex spatial context. In this thesis, we first propose a coarse-to-fine method (**RandomPatch (Chapter 5)**) for high-resolution

3D images, the two-stage method can effectively modeling hierarchical context by high- and low-res models. We also create an approach (**PredictivePhenotyping (Chapter 6)**) for taking both image and EHR data as input, which can modeling the phenotype information as the same time. After the storm of the transformer model, we investigate the 3D perspective of Vision Transformer [88] models in spatial medical image segmentation. We propose the **UNETR (Chapter 7)** model for CT and MRI images. Later, to benefit from the scalability and efficiency of transformer models, we propose the self-supervised pre-training technique with a new generic medical image segmentation method (**Self-Supervised Swin UNETR (Chapter 8)**). Furthermore, we focus on evaluating the scaling and efficient properties of transformer blocks in hybrid medical image segmentation models. Toward more efficient transformer encoder, we propose to use the 3D aggregation blocks within the encoder, named **UNEST (Chapter 9)**, the one achieves the challenging segmentation task of renal sub-structures and whole brain segmentation. We show that scaling up the network complexity can significantly benefit from the large-scale data pre-training and for low-data regime training in down-stream tasks. In summary, all these proposed segmentation methods tackle the challenges of hierarchical contexts in the spatial medical images. We achieve state-of-the-art performance for many segmentation benchmark datasets.

After accessing the effective tools for quantitative measurements, we focus on studying the abdominal tissue mapping. We build the first CT image based atlas templates on the queried population data. Using the affine and deformable registration, we construct the framework of **normal kidney atlas**), **kideny sub-structure atlas**, and **pancreas atlas (Chapter 10)** with CT images. Under HubMAP project, we investigate the stability of the constructed average template for each organ, and validate its usage for surface, tissue mapping.

We extensively study and validate the quantitative tools for clinical-oriented problems. First, we evaluate whether the automatic segmentation method is stable for segmenting splenomegaly data (**Splenomegaly Validation (Chapter 11)**). The performance and metrics on the enlarged spleens are sufficient to quantify and provide useful biomarkers for clinical workflows. In addition, we investigate five different types of EMR features to predict the onset of T2DM (**T2DM Prediction (Chapter 11)**). We show that each contextual feature from patients' clinical history improves the prediction of onset T2DM. Next, we construct a deep neural network for encoding pancreas CT slices for the T2DM onset prediction.

In general, we worked on data- and human-centric artificial intelligence systems for medical image analysis and healthcare. We lead research on efficient and robust deep learning for either public and real clinical data. We enable innovative methodologies applicable for real clinical problems.

12.2 Visions Beyond Medical Image Analysis

Our research mission is to explore leading technologies for healthcare scenarios. The science is desired to help clinicians, radiologist, and researchers for their daily work, or to use quantitative tools exploring new findings. Our vision of a successful artificial intelligence system in healthcare is not to replace individuals, but to create new approaches. We see current deep learning techniques are insufficient to achieve robustness and generalizability, making it inapplicable to real clinical workflow. We propose some technical perspectives for developing medical AI systems.

Vision 1: Efficient Models. Current deep learning models can only modeling small well-distributed data. With the increase of training model complexity, medical AI systems require efficient and scalable agent for learning useful contexts.

Vision 2: Multi-Modal Data. Clinical data from multiple modalities can convey multi-view context of a target. Different information modalities, such as image, reports, standard notes, ICD, CPT codes, and medicine history can compensate each other as an entry to biomedical informatics. The ability of joint modeling multi-modalities data can empower models with rich representation contexts.

Vision 3: Collaborated Learning. To increase and diversify the available data for medical AI systems. And instead of training isolated machine learning algorithms within a data center, Federated learning (FL) can lead to better leverage of multiple sites participating. The de-centralized training strategy brought by FL is effective for learning from data streams and adapting clients' models via transfer learning.

Vision 4: Life-Long Learning. The data-driven scene classification is an essential technique for image-guided medicine. We consider the life-long classification problem as the model is required to learn multiple concepts (categories) in sequential order through a stream of new training data points. In addition, it has been known in various clinical applications, that multiple biomarkers can be discovered for the same disease progressions over time. New explicit data signals are consistently evolving to impact existing tasks. Currently, transfer learning or fine-tuning ML sys-

tems without catastrophic forgetting is the key target for lifelong classification problems. Models for well-controlled small datasets and learning offline by connectionist models are suffering from interference and the loss of knowledge, which leads to the limited representation learning ability of deep neural networks. Thus, it is necessary to develop scalable networks that are robust to massive spatio-temporal information when the new instances are observed.

12.3 Future Works

In this section, we propose to explore future directions based on the current contributed works.

12.3.1 Scaling Transformer Medical Segmentation Models for Large-Scale Data

The self-attention-based transformers have been endorsed as prodigious encoders in modeling volumetric medical image representations. In comparison to the explicit and well-adopted convolutional neural networks (CNNs), transformer models re-designs the learning paradigm in a sequence-to-sequence mechanism, exhibiting larger receptive field for global context. However, the optimal scaling of transformer models as efficient encoder for 3D medical image segmentation remains unclear. We will focus on evaluating data efficiency and transformer scales for 3D medical images. We will investigate the efficient design of hierarchical transformers, the spatial block aggregation mechanism are employed for the transformer encoder. In addition, the scaling properties of the transformer model for medical image analysis remains unclear. We will study the properties of scaling up models for whole brain segmentation. Using the advantages of large GPU hardware, the scalability and robustness can be evaluated with large size of training data and low-data regime.

12.3.2 Understanding Imaging Biomarks for Type 1 Diabetes with CT and MRI images

Type I and II diabetes mellitus are common and significant chronic disease with both inherent and environmental causes. Diabetes is characterized by obesity with attendant risk factors including hyperglycemia, hypertension, and hyperglycemia stemming from insulin resistance. Potential markers of T1D and T2D include the aforementioned risk factors as well as regional obesity and pancreas changes which can be learned from patients' imaging and diagnostic history. Clinical framing of these variables relative to T2DM are complex through multiple risk factors, e.g., body mass index (BMI), pancreas tissue volume, visceral/subcutaneous fat distribution, and glucose tests. Previous

works have shown these hand-crafted features can be used to classify the presence. We will continue to study the quantitative biomarkers for metabolic syndrome related disease. Collaborating with UT Austin, we will continue to study the pancreas tissue pattern with MRI images.

12.3.3 Learning Continuously from Incremental Data and Organ Segmentation

Computation infrastructures are leading operations in the real world medical image understanding to increasingly continuous information instead of isolated data concepts and distributions. For instance, clinicians or medical agents interacting with the experience benefit from progressive knowledge over life-long time spans. The capability to increasingly learn new knowledge and retain progressively learned information named lifelong artificial intelligence (AI) system is attracting great application in image perceptions. Following the studies of multi-modalities data for CT and MRI abdominal data, we will investigate whether a model can learn different data incrementally (e.g., Learn CT images at first stage, then learn MRI images and pancreas label. The final model can take either CT or MRI image as input and segment pancreas.).

12.3.4 Collaborative Learning with Federated Clients for Healthcare

Following the study of learning multi-institute data, we will extend the collaborative learning with federated learning techniques. Healthcare data providers and clients can focus on training their own data, then the model can be distributed for inference to other without sharing patient data. The distributed aggregation servers will send the consensus model to all collaborating centers for fine-tuning models. This learning paradigm can well protect the data privacy while exploiting the multi-institute data.

References

- [1] F. Haghghi, M. R. H. Taher, Z. Zhou, M. B. Gotway, and J. Liang, "Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning," *IEEE transactions on medical imaging*, 2021.
- [2] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models genesis," *Medical image analysis*, vol. 67, p. 101840, 2021.
- [3] A. Tefferi, F. Cervantes, R. Mesa, F. Passamonti, S. Verstovsek, A. M. Vannucchi, J. Gotlib, B. Dupriez, A. Pardanani, and C. Harrison, "Revised response criteria for myelofibrosis: international working group-myeloproliferative neoplasms research and treatment (iwg-mrt) and european leukemianet (eln) consensus report," *Blood, The Journal of the American Society of Hematology*, vol. 122, no. 8, pp. 1395–1398, 2013.
- [4] Y. Tang, Y. Huo, Y. Xiong, H. Moon, A. Assad, T. K. Moyo, M. R. Savona, R. Abramson, and B. A. Landman, "Improving splenomegaly segmentation by learning from heterogeneous multi-source labels," in *Medical Imaging 2019: Image Processing*, vol. 10949, p. 1094908, International Society for Optics and Photonics.
- [5] S. B. Heymsfield, T. FULENWIDER, B. Nordlinger, R. Barlow, P. Sones, and M. Kutner, "Accurate measurement of liver, kidney, and spleen volume and mass by computerized axial tomography," *Annals of internal medicine*, vol. 90, no. 2, pp. 185–187, 1979.
- [6] J. Czernin, M. R. Benz, and M. S. Allen-Auerbach, "Pet/ct imaging: The incremental value of assessing the glucose metabolic phenotype and the structure of cancers in a single examination," *European journal of radiology*, vol. 73, no. 3, pp. 470–480, 2010.
- [7] E. Kulama, "Scanning protocols for multislice ct scanners," *The British journal of radiology*, vol. 77, no. suppl_1, pp. S2–S9, 2004.
- [8] D. J. Brenner and E. J. Hall, "Computed tomography—an increasing source of radiation exposure," *New England journal of medicine*, vol. 357, no. 22, pp. 2277–2284, 2007.
- [9] H. R. Roth, A. Farag, E. B. Turkbey, L. Lu, J. Liu, and R. M. Summers, "Data from pancreas-ct," *The cancer imaging archive*, 2016.
- [10] A. Muhi, T. Ichikawa, U. Motosugi, H. Sou, H. Nakajima, K. Sano, M. Sano, S. Kato, T. Kitamura, and Z. Fatima, "Diagnosis of colorectal hepatic metastases: comparison of contrast-enhanced ct, contrast-enhanced us, superparamagnetic iron oxide-enhanced mri, and gadoteric acid-enhanced mri," *Journal of Magnetic Resonance Imaging*, vol. 34, no. 2, pp. 326–335, 2011.
- [11] G. Tognini, F. Ferrozzi, D. Bova, P. Bini, and M. Zompatori, "Diabetes mellitus: Ct findings of unusual complications related to the disease: a pictorial essay," *Clinical imaging*, vol. 27, no. 5, pp. 325–329, 2003.
- [12] B. Dussol, J. Moussi-Frances, S. Morange, C. Somma-Delpero, O. Mundler, and Y. Berland, "A randomized trial of furosemide vs hydrochlorothiazide in patients with chronic renal failure and hypertension," *Nephrology Dialysis Transplantation*, vol. 20, no. 2, pp. 349–353, 2005.

- [13] A. B. Chapman, J. E. Bost, V. E. Torres, L. Guay-Woodford, K. T. Bae, D. Landsittel, J. Li, B. F. King, D. Martin, and L. H. Wetzel, “Kidney volume and functional outcomes in autosomal dominant polycystic kidney disease,” *Clinical Journal of the American Society of Nephrology*, vol. 7, no. 3, pp. 479–486, 2012.
- [14] T. M. Buzug, *Computed tomography*, pp. 311–342. Springer, 2011.
- [15] A. B. Rosenkrantz, M. Mendiratta-Lala, B. J. Bartholmai, D. Ganeshan, R. G. Abramson, K. R. Burton, J. Y. John-Paul, E. M. Scalzetti, T. E. Yankeelov, and R. M. Subramaniam, “Clinical utility of quantitative imaging,” *Academic radiology*, vol. 22, no. 1, pp. 33–49, 2015.
- [16] R. G. Abramson, K. R. Burton, J. Y. John-Paul, E. M. Scalzetti, T. E. Yankeelov, A. B. Rosenkrantz, M. Mendiratta-Lala, B. J. Bartholmai, D. Ganeshan, and L. Lenchik, “Methods and challenges in quantitative imaging biomarker development,” *Academic radiology*, vol. 22, no. 1, pp. 25–32, 2015.
- [17] K. Goda, E. Sasaki, K. Nagata, M. Fukai, N. Ohsawa, and T. Hahafusa, “Pancreatic volume in type 1 and type 2 diabetes mellitus,” *Acta diabetologica*, vol. 38, no. 3, pp. 145–149, 2001.
- [18] S. W. van den Dool, M. N. Wasser, J. W. de Fijter, J. Hoekstra, and R. J. van der Geest, “Functional renal volume: quantitative analysis at gadolinium-enhanced mr angiography—feasibility study in healthy potential kidney donors,” *Radiology*, vol. 236, no. 1, pp. 189–195, 2005.
- [19] Z. Xu, R. P. Burke, C. P. Lee, R. B. Baucom, B. K. Poulouse, R. G. Abramson, and B. A. Landman, “Efficient multi-atlas abdominal segmentation on clinically acquired ct with simple context learning,” *Medical image analysis*, vol. 24, no. 1, pp. 18–27, 2015.
- [20] M. J. McAuliffe, F. M. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. L. Trus, “Medical image processing, analysis and visualization in clinical research,” in *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001*, pp. 381–386, IEEE.
- [21] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research,” *Computational and structural biotechnology journal*, vol. 15, pp. 104–116, 2017.
- [22] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
- [23] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778, 2016.
- [25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 386–397, 2020.
- [26] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 234–241, Springer, 2015.

- [27] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.
- [28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.
- [29] D. Shen, G. Wu, and H.-I. Suk, “Deep learning in medical image analysis,” *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.
- [30] J. Hsieh, “Computed tomography: principles, design, artifacts, and recent advances,” SPIE Bellingham, WA.
- [31] K. Yan, L. Lu, and R. M. Summers, “Unsupervised body part regression via spatially self-ordering convolutional neural networks,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 1022–1025, IEEE.
- [32] X. Xu, F. Zhou, B. Liu, D. Fu, and X. Bai, “Efficient multiple organ localization in ct image using 3d region proposal network,” *IEEE transactions on medical imaging*, 2019.
- [33] K. Yan, X. Wang, L. Lu, and R. M. Summers, “Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning,” *Journal of Medical Imaging*, vol. 5, no. 3, p. 036501, 2018.
- [34] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, *et al.*, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic),” in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172, IEEE, 2018.
- [35] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, “3d deeply supervised network for automatic liver segmentation from ct volumes,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 149–157, Springer.
- [36] K. T. Bae, “Intravenous contrast medium administration and scan timing at ct: considerations and approaches,” *Radiology*, vol. 256, no. 1, pp. 32–61, 2010.
- [37] K. Awai and S. Hori, “Effect of contrast injection protocol with dose tailored to patient weight and fixed injection duration on aortic and hepatic enhancement at multidetector-row helical ct,” *European radiology*, vol. 13, no. 9, pp. 2155–2160, 2003.
- [38] K. Mitsuzaki, Y. Yamashita, I. Ogata, T. Nishiharu, J. Urata, and M. Takahashi, “Multiple-phase helical ct of the liver for detecting small hepatomas in patients with liver cirrhosis: contrast-injection protocol and optimal timing,” *AJR. American journal of roentgenology*, vol. 167, no. 3, pp. 753–757, 1996.
- [39] J. Oliver 3rd, R. Baron, M. Federle, and H. Rockette Jr, “Detecting hepatocellular carcinoma: value of unenhanced or arterial phase ct imaging or both used in conjunction with conventional portal venous phase contrast-enhanced ct imaging,” *AJR. American journal of roentgenology*, vol. 167, no. 1, pp. 71–77, 1996.

- [40] Y. Tang, H. H. Lee, Y. Xu, O. Tang, Y. Chen, D. Gao, S. Han, R. Gao, C. Bermudez, and M. R. Savona, "Contrast phase classification with a generative adversarial network," *arXiv preprint arXiv:1911.06395*, 2019.
- [41] B. Zhou, A. P. Harrison, J. Yao, C.-T. Cheng, J. Xiao, C.-H. Liao, and L. Lu, *CT Data Curation for Liver Patients: Phase Recognition in Dynamic Contrast-Enhanced CT*, pp. 139–147. Springer, 2019.
- [42] H. R. Roth, C. Shen, H. Oda, T. Sugino, M. Oda, Y. Hayashi, K. Misawa, and K. Mori, "A multi-scale pyramid of 3d fully convolutional networks for abdominal multi-organ segmentation," in *International conference on medical image computing and computer-assisted intervention*, 2018.
- [43] Y. Tang, R. Gao, H. H. Lee, S. Han, Y. Chen, D. Gao, V. Nath, C. Bermudez, M. R. Savona, R. G. Abramson, *et al.*, "High-resolution 3d abdominal segmentation with random patch network fusion," *Medical Image Analysis*, vol. 69, p. 101894, 2021.
- [44] Y. Zhou, Z. Li, S. Bai, C. Wang, X. Chen, M. Han, E. Fishman, and A. L. Yuille, "Prior-aware neural network for partially-supervised multi-organ segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [45] A. S. Bezerra, G. D'Ippolito, S. Faintuch, J. Szejnfeld, and M. Ahmed, "Determination of splenomegaly by ct: is there a place for a single measurement?," *American Journal of Roentgenology*, vol. 184, no. 5, pp. 1510–1513, 2005.
- [46] P. A. McCormick and K. M. Murphy, "Splenomegaly, hypersplenism and coagulation abnormalities in liver disease," *Best Practice Research Clinical Gastroenterology*, vol. 14, no. 6, pp. 1009–1031, 2000.
- [47] A. Woodruff, "Mechanisms involved in anaemia associated with infection and splenomegaly in the tropics," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 67, no. 3, pp. 313–25, 1973.
- [48] B. Klein, M. Stein, A. Kuten, M. Steiner, D. Barshalom, E. Robinson, and D. Gal, "Splenomegaly and solitary spleen metastasis in solid tumors," *Cancer*, vol. 60, no. 1, pp. 100–102, 1987.
- [49] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE transactions on medical imaging*, vol. 32, no. 9, pp. 1723–1730, 2013.
- [50] Z. Xu, B. Li, S. Panda, A. J. Asman, K. L. Merkle, P. L. Shanahan, R. G. Abramson, and B. A. Landman, "Shape-constrained multi-atlas segmentation of spleen in ct," in *Medical Imaging 2014: Image Processing*, vol. 9034, p. 903446, International Society for Optics and Photonics.
- [51] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

- [52] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, “Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy,” *Neuroimage*, vol. 46, no. 3, pp. 726–738, 2009.
- [53] A. J. Asman and B. A. Landman, “Non-local statistical label fusion for multi-atlas segmentation,” *Medical image analysis*, vol. 17, no. 2, pp. 194–208, 2013.
- [54] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich, “Multi-atlas segmentation with joint label fusion,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 3, pp. 611–623, 2012.
- [55] Y. Huo, Z. Xu, Y. Xiong, K. Aboud, P. Parvathaneni, S. B. NeuroImage, and u. 2019, “3d whole brain segmentation using spatially localized atlas network tiles,” *Elsevier*.
- [56] W. Bai, W. Shi, D. P. O’regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters, and D. Rueckert, “A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac mr images,” *IEEE transactions on medical imaging*, vol. 32, no. 7, pp. 1302–1315, 2013.
- [57] Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3d u-net: learning dense volumetric segmentation from sparse annotation,” in *International conference on medical image computing and computer-assisted intervention*, pp. 424–432, Springer.
- [58] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.
- [59] M. Lai, “Deep learning for medical image segmentation,” *arXiv preprint arXiv:1505.02000*, 2015.
- [60] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, “H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes,” *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.
- [61] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE.
- [62] H. R. Roth, H. Oda, Y. Hayashi, M. Oda, N. Shimizu, M. Fujiwara, K. Misawa, and K. Mori, “Hierarchical 3d fully convolutional networks for multi-organ segmentation,” *arXiv preprint arXiv:1704.06382*, 2017.
- [63] Z. Zhu, Y. Xia, W. Shen, E. Fishman, and A. Yuille, “A 3d coarse-to-fine framework for volumetric medical image segmentation,” in *2018 International Conference on 3D Vision (3DV)*, pp. 682–690, IEEE.
- [64] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv preprint arXiv:1706.05587*, 2017.
- [65] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer.

- [66] H. R. Roth, H. Oda, Y. Hayashi, M. Oda, N. Shimizu, M. Fujiwara, K. Misawa, and K. Mori, “Hierarchical 3d fully convolutional networks for multi-organ segmentation,” *arXiv preprint arXiv:1704.06382*, 2017.
- [67] H. Kim, J. Jung, J. Kim, B. Cho, J. Kwak, J. Y. Jang, S.-w. Lee, J.-G. Lee, and S. M. Yoon, “Abdominal multi-organ auto-segmentation using 3d-patch-based deep convolutional neural network,” *Scientific Reports*, vol. 10, no. 1, pp. 1–9, 2020.
- [68] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” 5 2015.
- [69] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [70] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *NIPS*, 2012.
- [71] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.
- [72] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, “Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge,” in *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.
- [73] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *European conference on computer vision*, pp. 649–666, Springer.
- [74] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.
- [75] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *European conference on computer vision*, 2016.
- [76] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [77] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, “Models genesis,” *Medical image analysis*, vol. 67, p. 101840, 2021.
- [78] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. Van Den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, and B. Geurts, “Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge,” *Medical image analysis*, vol. 42, pp. 1–13, 2017.
- [79] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *Journal of machine learning research*, vol. 11, no. 12, 2010.
- [80] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149.

- [81] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, “Self-supervised learning for medical image analysis using image context restoration,” *Medical image analysis*, vol. 58, p. 101539, 2019.
- [82] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, “Self-supervised feature learning for 3d medical images by playing a rubik’s cube,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 420–428, Springer.
- [83] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*, 2020.
- [84] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, pp. 5998–6008.
- [85] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [86] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [87] Y.-S. Chuang, C.-L. Liu, and H.-Y. Lee, “Speechbert: Cross-modal pre-trained language model for end-to-end spoken question answering,” 2019.
- [88] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2020.
- [89] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European Conference on Computer Vision*, 2020.
- [90] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, *et al.*, “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” *arXiv preprint arXiv:2012.15840*, 2020.
- [91] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, “End-to-end dense video captioning with masked transformer,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8739–8748.
- [92] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, “Unetr: Transformers for 3d medical image segmentation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, 2022.
- [93] Y. Xie, J. Zhang, C. Shen, and Y. Xia, “Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation,” *International conference on medical image computing and computer-assisted intervention*, 2021.
- [94] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Tran-sunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.

- [95] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care,” *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.
- [96] E. R. Pearson, “Type 2 diabetes: a multifaceted disease,” *Diabetologia*, vol. 62, no. 7, pp. 1107–1112, 2019.
- [97] X. Zhang, J. Chou, J. Liang, C. Xiao, Y. Zhao, H. Sarva, C. Henchcliffe, and F. Wang, “Data-driven subtyping of parkinson’s disease using longitudinal clinical records: a cohort study,” *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.
- [98] J. Virostko, M. Hilmes, K. Eitel, D. J. Moore, and A. C. Powers, “Use of the electronic medical record to assess pancreas size in type 1 diabetes,” *PloS one*, vol. 11, no. 7, 2016.
- [99] A. N. Kho, M. G. Hayes, L. Rasmussen-Torvik, J. A. Pacheco, W. K. Thompson, L. L. Armstrong, J. C. Denny, P. L. Peissig, A. W. Miller, and W.-Q. Wei, “Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study,” *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 212–218, 2012.
- [100] S. A. Dugger, A. Platt, and D. B. Goldstein, “Drug development in the era of precision medicine,” *Nature reviews Drug discovery*, vol. 17, no. 3, pp. 183–196, 2018.
- [101] C. Lee and M. Van Der Schaar, “Temporal phenotyping using deep predictive clustering of disease progression,” in *International Conference on Machine Learning*, pp. 5767–5777, PMLR.
- [102] A. Rusanov, P. V. Prado, and C. Weng, “Unsupervised time-series clustering over lab data for automatic identification of uncontrolled diabetes,” in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 72–80, IEEE.
- [103] E. Kawaler, A. Cobian, P. Peissig, D. Cross, S. Yale, and M. Craven, “Learning to predict post-hospitalization vte risk from ehr data,” in *AMIA annual symposium proceedings*, vol. 2012, p. 436, American Medical Informatics Association.
- [104] D. Zhao and C. Weng, “Combining pubmed knowledge and ehr data to develop a weighted bayesian network for pancreatic cancer prediction,” *Journal of biomedical informatics*, vol. 44, no. 5, pp. 859–868, 2011.
- [105] D. Chen, S. Liu, P. Kingsbury, S. Sohn, C. B. Storlie, E. B. Habermann, J. M. Naessens, D. W. Larson, and H. Liu, “Deep learning and alternative learning strategies for retrospective real-world clinical data,” *NPJ digital medicine*, vol. 2, no. 1, pp. 1–5, 2019.
- [106] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, “Patient subtyping via time-aware lstm networks,” in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 65–74.
- [107] G. E. Hinton, “Boltzmann machine,” *Scholarpedia*, vol. 2, no. 5, p. 1668, 2007.
- [108] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

- [109] Y. Mroueh, E. Marcheret, and V. Goel, “Deep multimodal learning for audio-visual speech recognition,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2130–2134, IEEE.
- [110] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.
- [111] Z. Zhang, L. Yang, and Y. Zheng, “Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9242–9251.
- [112] Z. Zhang, P. Chen, M. Sapkota, and L. Yang, “Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 320–328, Springer.
- [113] P. Zhang, F. Wang, and Y. Zheng, “Self supervised deep representation learning for fine-grained body part recognition,” in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 578–582, IEEE.
- [114] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*, vol. 589. John Wiley sons, 2005.
- [115] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3d u-net: learning dense volumetric segmentation from sparse annotation,” in *International conference on medical image computing and computer-assisted intervention*, 2016.
- [116] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [117] P. J. Rousseeuw, “Least median of squares regression,” *Journal of the American statistical association*, vol. 79, no. 388, pp. 871–880, 1984.
- [118] Y. Liu and M. S. Lew, “Learning relaxed deep supervision for better edge detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 231–240.
- [119] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, “3d deeply supervised network for automatic liver segmentation from ct volumes,” in *International conference on medical image computing and computer-assisted intervention*, 2016.
- [120] S. Xie and Z. Tu, “Holistically-nested edge detection,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403.
- [121] Y. Tang, R. Gao, H. H. Lee, Q. S. Wells, A. Spann, J. G. Terry, J. J. Carr, Y. Huo, S. Bao, and B. A. Landman, “Prediction of type ii diabetes onset with computed tomography and electronic medical records,” in *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures*, pp. 13–23, Springer, 2020.

- [122] B. Glocker, J. Feulner, A. Criminisi, D. R. Haynor, and E. Konukoglu, “Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 590–598, Springer, 2012.
- [123] Y. Tang, R. Gao, H. H. Lee, Z. Xu, B. V. Savoie, S. Bao, Y. Huo, A. B. Fogo, R. Harris, M. P. de Caestecker, *et al.*, “Renal cortex, medulla and pelvicaliceal system segmentation on arterial phase ct images with random patch-based networks,” in *Medical Imaging 2021: Image Processing*, vol. 11596, p. 115961D, International Society for Optics and Photonics, 2021.
- [124] K. Yan, J. Cai, D. Jin, S. Miao, A. P. Harrison, D. Guo, Y. Tang, J. Xiao, J. Lu, and L. Lu, “Self-supervised learning of pixel-wise anatomical embeddings in radiological images,” *arXiv preprint arXiv:2012.02383*, 2020.
- [125] Y. Tang, R. Gao, S. Han, Y. Chen, D. Gao, V. Nath, C. Bermudez, M. R. Savona, S. Bao, I. Lyu, *et al.*, “Body part regression with self-supervision,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1499–1507, 2021.
- [126] X. Chen, S. Xie, and K. He, “An empirical study of training self-supervised vision transformers,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [127] Y. Tang, D. Yang, W. Li, H. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, “Self-supervised pre-training of swin transformers for 3d medical image analysis,” *arXiv preprint arXiv:2111.14791*, 2021.
- [128] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- [129] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, “Mixmatch: A holistic approach to semi-supervised learning,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [130] R. L. Baron, J. Oliver 3rd, G. Dodd 3rd, M. Nalesnik, B. L. Holbert, and B. Carr, “Hepatocellular carcinoma: evaluation with biphasic, contrast-enhanced, helical ct,” *Radiology*, vol. 199, no. 2, pp. 505–511, 1996.
- [131] Y. Huo, Y. Tang, Y. Chen, D. Gao, S. Han, S. Bao, S. De, J. G. Terry, J. J. Carr, and R. G. Abramson, “Stochastic tissue window normalization of deep learning on computed tomography,” *Journal of Medical Imaging*, vol. 6, no. 4, p. 044005, 2019.
- [132] T. Kanzaki and H. Sakagami, “Late phase allergic reaction to a ct contrast medium (iotrolan),” *The Journal of dermatology*, vol. 18, no. 9, pp. 528–531, 1991.
- [133] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
- [134] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” 9 2014.
- [135] Z. Zhang and M. Sabuncu, “Generalized cross entropy loss for training deep neural networks with noisy labels,” in *Advances in neural information processing systems*, pp. 8778–8788.

- [136] Y. Pu, Z. Gan, R. Henaio, X. Yuan, C. Li, A. Stevens, and L. Carin, “Variational auto-encoder for deep learning of images, labels and captions,” in *Advances in neural information processing systems*, pp. 2352–2360.
- [137] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.
- [138] L. Tran, X. Yin, and X. Liu, “Disentangled representation learning gan for pose-invariant face recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1415–1424.
- [139] Z. Zhang, L. Yang, and Y. Zheng, “Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9242–9251, 2018.
- [140] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel, “Infogan: Interpretable representation learning by information maximizing generative adversarial nets,” in *Advances in neural information processing systems*, pp. 2172–2180.
- [141] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, “Synseg-net: Synthetic segmentation without target modality ground truth,” *IEEE transactions on medical imaging*, vol. 38, no. 4, pp. 1016–1025, 2018.
- [142] H. Liao, Y. Tang, G. Funka-Lea, J. Luo, and S. K. Zhou, “More knowledge is better: Cross-modality volume completion and 3d+ 2d segmentation for intracardiac echocardiography contouring,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 535–543, Springer.
- [143] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, “StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8789–8797, IEEE Computer Society, 12 2018.
- [144] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Gaugan: semantic image synthesis with spatially adaptive normalization,” in *ACM SIGGRAPH 2019 Real-Time Live!*, p. 2, ACM.
- [145] T. Okada, M. G. Linguraru, Y. Yoshida, M. Hori, R. M. Summers, Y.-W. Chen, N. Tomiyama, and Y. Sato, “Abdominal multi-organ segmentation of ct images based on hierarchical spatial modeling of organ interrelations,” in *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging*, pp. 173–180, Springer.
- [146] J. M. Lacomis, R. L. Baron, J. Oliver 3rd, M. A. Nalesnik, and M. P. Federle, “Cholangiocarcinoma: delayed ct contrast enhancement patterns,” *Radiology*, vol. 203, no. 1, pp. 98–104, 1997.
- [147] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *NIPS*, vol. 3, pp. 2672–2680, 2014.
- [148] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” 11 2016.

- [149] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” in *ICCV*, pp. 2242–2251, Institute of Electrical and Electronics Engineers Inc., 12 2017.
- [150] W. He, B. Li, and D. Song, “Decision boundary analysis of adversarial examples,” 2018.
- [151] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” *arXiv preprint arXiv:1809.11096*, 2018.
- [152] J. Liu, Y. Huo, Z. Xu, A. Assad, R. G. Abramson, and B. A. Landman, “Multi-atlas spleen segmentation on ct using adaptive context learning,” in *Medical Imaging 2017: Image Processing*, vol. 10133, p. 1013309, International Society for Optics and Photonics.
- [153] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.
- [154] Y. Xu, O. Tang, Y. Tang, H. H. Lee, Y. Chen, D. Gao, S. Han, R. Gao, M. R. Savona, and R. G. Abramson, “Validation and optimization of multi-organ segmentation on clinical imaging archives,” *arXiv preprint arXiv:2002.04102*, 2020.
- [155] P. Moeskops, J. M. Wolterink, B. H. van der Velden, K. G. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, “Deep learning for multi-task medical image segmentation in multiple modalities,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 478–486, Springer.
- [156] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 fourth international conference on 3D vision (3DV)*, 2016.
- [157] A. de Brebisson and G. Montana, “Deep neural networks for anatomical brain segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–28.
- [158] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, “Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation,” *NeuroImage*, vol. 54, no. 2, pp. 940–954, 2011.
- [159] D. Zhang, Q. Guo, G. Wu, and D. Shen, “Sparse patch-based label fusion for multi-atlas segmentation,” in *International Workshop on Multimodal Brain Image Analysis*, pp. 94–102, Springer.
- [160] W. Yang, R. Gao, Y. Xu, X. Sun, and Q. Liao, “Discriminative patch-based sparse representation for face recognition,” in *ICSPCC 2016 - IEEE International Conference on Signal Processing, Communications and Computing, Conference Proceedings*, 2016.
- [161] Y. Huo, Z. Xu, Y. Xiong, K. Aboud, P. Parvathaneni, S. Bao, C. Bermudez, S. M. Resnick, L. E. Cutting, and B. A. Landman, “3d whole brain segmentation using spatially localized atlas network tiles,” *NeuroImage*, vol. 194, pp. 105–119, 2019.
- [162] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448.

- [163] G. Urban, M. Bendszus, F. Hamprecht, and J. Kleesiek, “Multi-modal brain tumor segmentation using deep convolutional neural networks,” *MICCAI BraTS (Brain Tumor Segmentation) Challenge. Proceedings, winning contribution*, pp. 31–35, 2014.
- [164] K. Kamnitsas, L. Chen, C. Ledig, D. Rueckert, and B. Glocker, “Multi-scale 3d convolutional neural networks for lesion segmentation in brain mri,” *Ischemic stroke lesion segmentation*, vol. 13, p. 46, 2015.
- [165] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818.
- [166] H. R. Roth, A. Farag, L. Lu, E. B. Turkbey, and R. M. Summers, “Deep convolutional networks for pancreas segmentation in ct imaging,” in *Medical Imaging 2015: Image Processing*, vol. 9413, p. 94131G, International Society for Optics and Photonics.
- [167] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” *arXiv preprint arXiv:1805.10180*, 2018.
- [168] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, “A fixed-point model for pancreas segmentation in abdominal ct scans,” in *International conference on medical image computing and computer-assisted intervention*, pp. 693–701, Springer.
- [169] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” in *International Conference on Information Processing in Medical Imaging*, pp. 146–157, Springer.
- [170] S. F. Eskildsen, P. Coupé, V. Fonov, J. V. Manjón, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Østergaard, D. L. Collins, and A. D. N. Initiative, “Beast: brain extraction based on nonlocal segmentation technique,” *NeuroImage*, vol. 59, no. 3, pp. 2362–2373, 2012.
- [171] A. J. Asman, Y. Huo, A. J. Plassard, and B. A. Landman, “Multi-atlas learner fusion: An efficient segmentation approach for large-scale data,” *Medical image analysis*, vol. 26, no. 1, pp. 82–91, 2015.
- [172] J. Ding, B. Chen, H. Liu, and M. Huang, “Convolutional neural network with data augmentation for sar target recognition,” *IEEE Geoscience and remote sensing letters*, vol. 13, no. 3, pp. 364–368, 2016.
- [173] I. Cheheb, N. Al-Maadeed, S. Al-Madeed, A. Bouridane, and R. Jiang, “Random sampling for patch-based face recognition,” in *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–5, IEEE.
- [174] L. Liang, C. Liu, Y.-Q. Xu, B. Guo, and H.-Y. Shum, “Real-time texture synthesis by patch-based sampling,” *ACM Transactions on Graphics (ToG)*, vol. 20, no. 3, pp. 127–150, 2001.
- [175] P. Coupé, B. Mansencal, M. Clément, R. Giraud, B. D. de Senneville, V.-T. Ta, V. Lepetit, and J. V. Manjon, “Assemblynet: A novel deep decision-making process for whole brain mri segmentation,” *arXiv preprint arXiv:1906.01862*, 2019.
- [176] H. R. Roth, C. Shen, H. Oda, T. Sugino, M. Oda, Y. Hayashi, K. Misawa, and K. Mori, “A multi-scale pyramid of 3d fully convolutional networks for abdominal multi-organ segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 417–425, Springer.

- [177] R. Olivier and C. Hanqiang, “Nearest neighbor value interpolation,” *arXiv preprint arXiv:1211.1768*, 2012.
- [178] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, “H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes,” *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.
- [179] H. R. Roth, H. Oda, X. Zhou, N. Shimizu, Y. Yang, Y. Hayashi, M. Oda, M. Fujiwara, K. Misawa, and K. Mori, “An application of cascaded 3d fully convolutional networks for medical image segmentation,” *Computerized Medical Imaging and Graphics*, vol. 66, pp. 90–99, 2018.
- [180] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” *arXiv:1802.02611*, 2018.
- [181] J. Chen, L. Yang, Y. Zhang, M. Alber, and D. Z. Chen, “Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3044–3052, 2016.
- [182] K. Goda, E. Sasaki, K. Nagata, M. Fukai, N. Ohsawa, and T. Hahafusa, “Pancreatic volume in type 1 und type 2 diabetes mellitus,” *Acta diabetologica*, vol. 38, no. 3, pp. 145–149, 2001.
- [183] H. R. Roth, A. Farag, L. Lu, E. B. Turkbey, and R. M. Summers, “Deep convolutional networks for pancreas segmentation in ct imaging,” in *Medical Imaging 2015: Image Processing*, vol. 9413, p. 94131G, International Society for Optics and Photonics, 2015.
- [184] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, “A fixed-point model for pancreas segmentation in abdominal ct scans,” in *MICCAI*, pp. 693–701, 2017.
- [185] Z. Zhu, Y. Xia, W. Shen, E. Fishman, and A. Yuille, “A 3d coarse-to-fine framework for volumetric medical image segmentation,” in *2018 International conference on 3D vision (3DV)*, pp. 682–690, IEEE, 2018.
- [186] C. N. Hales and D. J. Barker, “Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis,” *Diabetologia*, vol. 35, no. 7, pp. 595–601, 1992.
- [187] S. Mani, Y. Chen, T. Elasy, W. Clayton, and J. Denny, “Type 2 diabetes risk forecasting from emr data using machine learning,” in *AMIA annual symposium proceedings*, vol. 2012, p. 606, American Medical Informatics Association, 2012.
- [188] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali, “Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data,” *Medical care*, pp. 1130–1139, 2005.
- [189] J. A. Evans, “Electronic medical records system,” July 13 1999. US Patent 5,924,074.
- [190] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, “A machine learning-based framework to identify type 2 diabetes through electronic health records,” *International journal of medical informatics*, vol. 97, pp. 120–127, 2017.
- [191] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, “Patient subtyping via time-aware lstm networks,” in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 65–74, 2017.

- [192] A. Giannoula, A. Gutierrez-Sacristán, Á. Bravo, F. Sanz, and L. I. Furlong, “Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study,” *Scientific reports*, vol. 8, no. 1, pp. 1–14, 2018.
- [193] D. T. A. Luong and V. Chandola, “A k-means approach to clustering disease progressions,” in *2017 IEEE International conference on healthcare informatics (ICHI)*, pp. 268–274, IEEE, 2017.
- [194] C. Lee and M. Van Der Schaar, “Temporal phenotyping using deep predictive clustering of disease progression,” in *International Conference on Machine Learning*, pp. 5767–5777, PMLR, 2020.
- [195] J. Virostko, M. Hilmes, K. Eitel, D. J. Moore, and A. C. Powers, “Use of the electronic medical record to assess pancreas size in type 1 diabetes,” *PloS one*, vol. 11, no. 7, p. e0158825, 2016.
- [196] H. Roth, A. Farag, E. Turkbey, L. Lu, J. Liu, and R. Summers, “Data from pancreas-ct. the cancer imaging archive (2016).”
- [197] S. K. Zhou, H. Greenspan, and D. Shen, *Deep learning for medical image analysis*. Academic Press, 2017.
- [198] J. Xie, R. Girshick, and A. Farhadi, “Unsupervised deep embedding for clustering analysis,” in *International conference on machine learning*, pp. 478–487, PMLR, 2016.
- [199] I. Misra and L. v. d. Maaten, “Self-supervised learning of pretext-invariant representations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.
- [200] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.,” *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [201] H. Roth, L. Lu, A. Farag, H. Shin, J. Liu, E. Turkbey, and R. Summers, “DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation,” *MICCAI*, 2015.
- [202] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European Conference on Computer Vision*, pp. 213–229, Springer.
- [203] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” *arXiv preprint arXiv:2010.04159*, 2020.
- [204] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.
- [205] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021.
- [206] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, “Twins: Revisiting the design of spatial attention in vision transformers,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.

- [207] W. Xu, Y. Xu, T. Chang, and Z. Tu, “Co-scale conv-attentional image transformers,” *arXiv preprint arXiv:2104.06399*, 2021.
- [208] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.
- [209] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, “Medical transformer: Gated axial-attention for medical image segmentation,” *arXiv preprint arXiv:2102.10662*, 2021.
- [210] Y. Zhang, H. Liu, and Q. Hu, “Transfuse: Fusing transformers and cnns for medical image segmentation,” *arXiv preprint arXiv:2102.08005*, 2021.
- [211] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, *et al.*, “The medical segmentation decathlon,” *arXiv preprint arXiv:2106.05735*, 2021.
- [212] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, *et al.*, “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [213] B. Cheng, A. G. Schwing, and A. Kirillov, “Per-pixel classification is not all you need for semantic segmentation,” *arXiv preprint arXiv:2107.06278*, 2021.
- [214] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, “Unetr: Transformers for 3d medical image segmentation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, 2022.
- [215] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, “Scaling vision transformers,” *arXiv preprint arXiv:2106.04560*, 2021.
- [216] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, “Do vision transformers see like convolutional neural networks?,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [217] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [218] Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, “Self-supervised learning with swin transformers,” *arXiv preprint arXiv:2105.04553*, 2021.
- [219] S. Chen, K. Ma, and Y. Zheng, “Med3D: Transfer learning for 3D medical image analysis,” *arXiv:1904.00625*, 2019.
- [220] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding transfer learning for medical imaging,” in *Advances in neural information processing systems*, pp. 3347–3357, 2019.
- [221] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.

- [222] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” in *International Conference on Learning Representations*, 2018.
- [223] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.
- [224] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, “nnformer: Interleaved transformer for volumetric segmentation,” *arXiv preprint arXiv:2109.03201*, 2021.
- [225] J. Jose and P. Oza, “Medical transformer: gated axial-attention for medical image segmentation,” in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2021.
- [226] G. Xu, X. Wu, X. Zhang, and X. He, “Levit-unet: Make faster encoders with transformer for medical image segmentation,” *arXiv preprint arXiv:2107.08623*, 2021.
- [227] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017.
- [228] B. Singh, M. Najibi, and L. S. Davis, “Sniper: Efficient multi-scale training,” *Advances in Neural Information Processing Systems*, vol. 31, pp. 9310–9320, 2018.
- [229] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, “Video swin transformer,” *arXiv preprint arXiv:2106.13230*, 2021.
- [230] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, “Swin-unet: Unet-like pure transformer for medical image segmentation,” *arXiv preprint arXiv:2105.05537*, 2021.
- [231] A. Lin, B. Chen, J. Xu, Z. Zhang, and G. Lu, “Ds-transunet: Dual swin transformer u-net for medical image segmentation,” *arXiv preprint arXiv:2106.06716*, 2021.
- [232] S. Atito, M. Awais, and J. Kittler, “Sit: Self-supervised vision transformer,” *arXiv preprint arXiv:2104.03602*, 2021.
- [233] Z. Dai, B. Cai, Y. Lin, and J. Chen, “Up-detr: Unsupervised pre-training for object detection with transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [234] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “Swinir: Image restoration using swin transformer,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [235] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, “Self-supervised learning for medical image analysis using image context restoration,” *Medical image analysis*, vol. 58, p. 101539, 2019.
- [236] X. Wang, Z. Xu, L. Tam, D. Yang, and D. Xu, “Self-supervised image-text pre-training with mixed data in chest x-rays,” *arXiv preprint arXiv:2103.16022*, 2021.
- [237] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, *et al.*, “Big self-supervised models advance medical image classification,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

- [238] J. Zhu, Y. Li, Y. Hu, K. Ma, S. K. Zhou, and Y. Zheng, “Rubik’s cube+: A self-supervised feature learning framework for 3D medical image analysis,” *Medical Image Analysis*, p. 101746, 2020.
- [239] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, “3d self-supervised methods for medical imaging,” in *Advances in Neural Information Processing Systems*, 2020.
- [240] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” *arXiv preprint arXiv:1607.08022*, 2016.
- [241] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.
- [242] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, “Contrastive learning for unpaired image-to-image translation,” in *European Conference on Computer Vision*, 2020.
- [243] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *European Conference on Computer Vision*, 2018.
- [244] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations*, 2018.
- [245] S. Nikolov, S. Blackwell, A. Zverovitch, R. Mendes, M. Livne, J. De Fauw, Y. Patel, C. Meyer, H. Askham, B. Romera-Paredes, *et al.*, “Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy,” *arXiv preprint arXiv:1809.04430*, 2018.
- [246] Y. He, D. Yang, H. Roth, C. Zhao, and D. Xu, “Dints: Differentiable neural network topology search for 3d medical image segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [247] S. Kim, I. Kim, S. Lim, W. Baek, C. Kim, H. Cho, B. Yoon, and T. Kim, “Scalable neural architecture search for 3d medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019.
- [248] Q. Yu, D. Yang, H. Roth, Y. Bai, Y. Zhang, A. L. Yuille, and D. Xu, “C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [249] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. Van Den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, *et al.*, “Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge,” *Medical image analysis*, vol. 42, pp. 1–13, 2017.
- [250] S. Desai, A. Baghal, T. Wongsurawat, P. Jenjaroenpun, T. Powell, S. Al-Shukri, K. Gates, P. Farmer, M. Rutherford, G. Blake, *et al.*, “Chest imaging representing a covid-19 positive rural us population,” *Scientific data*, vol. 7, no. 1, pp. 1–6, 2020.

- [251] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, “The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans,” *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.
- [252] A. J. Grossberg, A. S. Mohamed, H. Elhalawani, W. C. Bennett, K. E. Smith, T. S. Nolan, B. Williams, S. Chamchod, J. Heukelom, M. E. Kantor, *et al.*, “Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy,” *Scientific data*, vol. 5, no. 1, pp. 1–10, 2018.
- [253] C. D. Johnson, M.-H. Chen, A. Y. Toledano, J. P. Heiken, A. Dachman, M. D. Kuo, C. O. Menias, B. Siewert, J. I. Cheema, R. G. Obregon, *et al.*, “Accuracy of ct colonography for detection of large adenomas and cancers,” *New England Journal of Medicine*, vol. 359, no. 12, pp. 1207–1217, 2008.
- [254] J.-B. Cordonnier, A. Loukas, and M. Jaggi, “On the relationship between self-attention and convolutional layers,” in *International Conference on Learning Representations*, 2019.
- [255] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” *arXiv preprint arXiv:2004.05150*, 2020.
- [256] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, “Transformer in transformer,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [257] V. S. Lee, H. Rusinek, M. E. Noz, P. Lee, M. Raghavan, and E. L. Kramer, “Dynamic three-dimensional mr renography for the measurement of single kidney function: initial experience,” *Radiology*, vol. 227, no. 1, pp. 289–294, 2003.
- [258] S. W. van den Dool, M. N. Wasser, J. W. de Fijter, J. Hoekstra, and R. J. van der Geest, “Functional renal volume: quantitative analysis at gadolinium-enhanced mr angiography—feasibility study in healthy potential kidney donors,” *Radiology*, vol. 236, no. 1, pp. 189–195, 2005.
- [259] D. V. Sahani, N. Rastogi, A. C. Greenfield, S. P. Kalva, D. Ko, S. Saini, G. Harris, and P. R. Mueller, “Multi-detector row ct in evaluation of 94 living renal donors by readers with varied experience,” *Radiology*, vol. 235, no. 3, pp. 905–910, 2005.
- [260] Z. Zhang, H. Zhang, L. Zhao, T. Chen, S. O. Arik, and T. Pfister, “Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding,” *arXiv preprint arXiv:2105.12723*, 2021.
- [261] W. Wang, C. Chen, M. Ding, J. Li, H. Yu, and S. Zha, “Transbts: Multimodal brain tumor segmentation using transformer,” *arXiv preprint arXiv:2103.04430*, 2021.
- [262] Y. Zhang, H. Liu, and Q. Hu, “Transfuse: Fusing transformers and cnns for medical image segmentation,” *arXiv preprint arXiv:2102.08005*, 2021.
- [263] Q. Jia and H. Shu, “Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation,” *arXiv preprint arXiv:2109.12271*, 2021.
- [264] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, “A volumetric transformer for accurate 3d tumor segmentation,” *arXiv preprint arXiv:2111.13300*, 2021.

- [265] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, “Medical transformer: Gated axial-attention for medical image segmentation,” *arXiv preprint arXiv:2102.10662*, 2021.
- [266] Y. Chang, H. Menghan, Z. Guangtao, and Z. Xiao-Ping, “Transclaw u-net: Claw u-net with transformers for medical image segmentation,” *arXiv preprint arXiv:2107.05188*, 2021.
- [267] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, “Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images,” *arXiv preprint arXiv:2201.01266*, 2022.
- [268] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 3–11, Springer, 2018.
- [269] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, *et al.*, “The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge,” *Medical image analysis*, vol. 67, p. 101821, 2021.
- [270] X. Chen, R. M. Summers, M. Cho, U. Bagci, and J. Yao, “An automatic method for renal cortex segmentation on ct images: evaluation on kidney donors,” *Academic radiology*, 2012.
- [271] D. Xiang, U. Bagci, C. Jin, F. Shi, W. Zhu, J. Yao, M. Sonka, and X. Chen, “Cortexpert: A model-based method for automatic renal cortex segmentation,” *Medical image analysis*, vol. 42, pp. 257–273, 2017.
- [272] C. Jin, F. Shi, D. Xiang, X. Jiang, B. Zhang, X. Wang, W. Zhu, E. Gao, and X. Chen, “3d fast automatic segmentation of kidney based on modified aam and random forest,” *IEEE transactions on medical imaging*, vol. 35, no. 6, pp. 1395–1407, 2016.
- [273] O. Rozenblatt-Rosen, M. J. Stubbington, A. Regev, and S. A. Teichmann, “The human cell atlas: from vision to reality,” *Nature*, vol. 550, no. 7677, pp. 451–453, 2017.
- [274] H. Consortium *et al.*, “The human body at cellular resolution: the nih human biomolecular atlas program,” *Nature*, vol. 574, no. 7777, p. 187, 2019.
- [275] M. P. Heinrich, M. Jenkinson, B. W. Papież, S. M. Brady, and J. A. Schnabel, “Towards real-time multimodal fusion for image-guided interventions using self-similarities,” in *International conference on medical image computing and computer-assisted intervention*, pp. 187–194, Springer, 2013.
- [276] M. P. Heinrich, O. Maier, and H. Handels, “Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities,” *VISCERAL Challenge@ ISBI*, vol. 1390, p. 27, 2015.
- [277] V. S. Lee, H. Rusinek, M. E. Noz, P. Lee, M. Raghavan, and E. L. Kramer, “Dynamic three-dimensional mr renography for the measurement of single kidney function: initial experience,” *Radiology*, vol. 227, no. 1, pp. 289–294, 2003.
- [278] A. Holden, A. Smith, P. Dukes, H. Pilmore, and M. Yasutomi, “Assessment of 100 live potential renal donors for laparoscopic nephrectomy with multi-detector row helical ct,” *Radiology*, vol. 237, no. 3, pp. 973–980, 2005.

- [279] E. J. Halpern, D. G. Mitchell, R. J. Wechsler, E. K. Outwater, M. J. Moritz, and G. A. Wilson, "Preoperative evaluation of living renal donors: comparison of ct angiography and mr angiography," *Radiology*, vol. 216, no. 2, pp. 434–439, 2000.
- [280] D. V. Sahani, N. Rastogi, A. C. Greenfield, S. P. Kalva, D. Ko, S. Saini, G. Harris, and P. R. Mueller, "Multi-detector row ct in evaluation of 94 living renal donors by readers with varied experience," *Radiology*, vol. 235, no. 3, pp. 905–910, 2005.
- [281] P. Jackson, N. Hardcastle, N. Dawe, T. Kron, M. S. Hofman, and R. J. Hicks, "Deep learning renal segmentation for fully automated radiation dose estimation in unsealed source therapy," *Frontiers in oncology*, vol. 8, p. 215, 2018.
- [282] C. Jin, F. Shi, D. Xiang, X. Jiang, B. Zhang, X. Wang, W. Zhu, E. Gao, and X. Chen, "3d fast automatic segmentation of kidney based on modified aam and random forest," *IEEE transactions on medical imaging*, vol. 35, no. 6, pp. 1395–1407, 2016.
- [283] D. Xiang, U. Bagci, C. Jin, F. Shi, W. Zhu, J. Yao, M. Sonka, and X. Chen, "Cortexpert: A model-based method for automatic renal cortex segmentation," *Medical image analysis*, vol. 42, pp. 257–273, 2017.
- [284] H. Shim, S. Chang, C. Tao, J. H. Wang, D. Kaya, and K. T. Bae, "Semiautomated segmentation of kidney from high-resolution multidetector computed tomography images using a graph-cuts technique," *Journal of computer assisted tomography*, vol. 33, no. 6, pp. 893–901, 2009.
- [285] X. Chen, R. M. Summers, M. Cho, U. Bagci, and J. Yao, "An automatic method for renal cortex segmentation on ct images: evaluation on kidney donors," *Academic radiology*, vol. 19, no. 5, pp. 562–570, 2012.
- [286] X. Li, X. Chen, J. Yao, X. Zhang, and J. Tian, "Renal cortex segmentation using optimal surface search with novel graph construction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 387–394, Springer.
- [287] S. Verstovsek, R. A. Mesa, J. Gotlib, R. S. Levy, V. Gupta, J. F. DiPersio, J. V. Catalano, M. Deininger, C. Miller, and R. T. Silver, "A double-blind, placebo-controlled trial of ruxolitinib for myelofibrosis," *New England Journal of Medicine*, vol. 366, no. 9, pp. 799–807, 2012.
- [288] A. M. Vannucchi, H. M. Kantarjian, J.-J. Kiladjan, J. Gotlib, F. Cervantes, R. A. Mesa, N. J. Sarlis, W. Peng, V. Sandor, and P. Gopalakrishna, "A pooled analysis of overall survival in comfort-i and comfort-ii, 2 randomized phase iii trials of ruxolitinib for the treatment of myelofibrosis," *haematologica*, vol. 100, no. 9, pp. 1139–1145, 2015.
- [289] Y. Tang, Y. Huo, Y. Xiong, H. Moon, A. Assad, T. K. Moyo, M. R. Savona, R. Abramson, and B. A. Landman, "Improving splenomegaly segmentation by learning from heterogeneous multi-source labels," in *Medical Imaging 2019: Image Processing*, vol. 10949, p. 1094908, International Society for Optics and Photonics, 2019.
- [290] Y. Huo, Z. Xu, S. Bao, C. Bermudez, A. J. Plassard, J. Liu, Y. Yao, A. Assad, R. G. Abramson, and B. A. Landman, "Splenomegaly segmentation using global convolutional kernels and conditional generative adversarial networks," in *Medical Imaging 2018: Image Processing*, vol. 10574, p. 1057409, International Society for Optics and Photonics.

- [291] R. G. Abramson and T. E. Yankeelov, “Imaging biomarkers and surrogate endpoints in oncology clinical trials,” in *Functional imaging in oncology*, pp. 29–42, Springer, 2014.
- [292] L. P. Clarke, R. J. Nordstrom, H. Zhang, P. Tandon, Y. Zhang, G. Redmond, K. Farahani, G. Kelloff, L. Henderson, and L. Shankar, “The quantitative imaging network: Nci’s historical perspective and planned goals,” *Translational oncology*, vol. 7, no. 1, p. 1, 2014.
- [293] A. J. Buckler, L. Bresolin, N. R. Dunnick, D. C. Sullivan, and Group, “A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging,” *Radiology*, vol. 258, no. 3, pp. 906–914, 2011.
- [294] S. H. Park and K. Han, “Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction,” *Radiology*, vol. 286, no. 3, pp. 800–809, 2018.
- [295] Y. Huo, J. Liu, Z. Xu, R. L. Harrigan, A. Assad, R. G. Abramson, and B. A. Landman, “Robust multicontrast mri spleen segmentation for splenomegaly using multi-atlas segmentation,” *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 2, pp. 336–343, 2017.
- [296] Z. Xu, A. L. Gertz, R. P. Burke, N. Bansal, H. Kang, B. A. Landman, and R. G. Abramson, “Improving spleen volume estimation via computer-assisted segmentation on clinically acquired ct scans,” *Academic radiology*, vol. 23, no. 10, pp. 1214–1220, 2016.
- [297] M. G. Linguraru, J. K. Sandberg, Z. Li, F. Shah, and R. M. Summers, “Automated segmentation and quantification of liver and spleen from ct images using normalized probabilistic atlases and enhancement estimation,” *Medical physics*, vol. 37, no. 2, pp. 771–783, 2010.
- [298] J. W. Lee, V. Devanarayan, Y. C. Barrett, R. Weiner, J. Allinson, S. Fountain, S. Keller, I. Weinryb, M. Green, and L. Duan, “Fit-for-purpose method development and validation for successful biomarker measurement,” *Pharmaceutical research*, vol. 23, no. 2, pp. 312–328, 2006.
- [299] D. J. Sargent, L. Rubinstein, L. Schwartz, J. Dancey, C. Gatsonis, L. Dodd, and L. Shankar, “Validation of novel imaging methodologies for use as cancer clinical trial end-points,” *European journal of cancer*, vol. 45, no. 2, pp. 290–299, 2009.
- [300] A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool,” *BMC Medical Imaging*, vol. 15, pp. 1–28, 8 2015.

Appendix A

Copyright from Publishers

A.1 Copyright from arXiv

Our publication work (Chapter 9) is under the license of CC BY-NC-ND. I, Yucheng Tang, am the creator and holder, retain ownership of the manuscript. A screenshot of copyright/License information are shown in A.1. Our Chapter 3 and 10 is under preparation for submission, whose newer version may be appearing in arXiv.

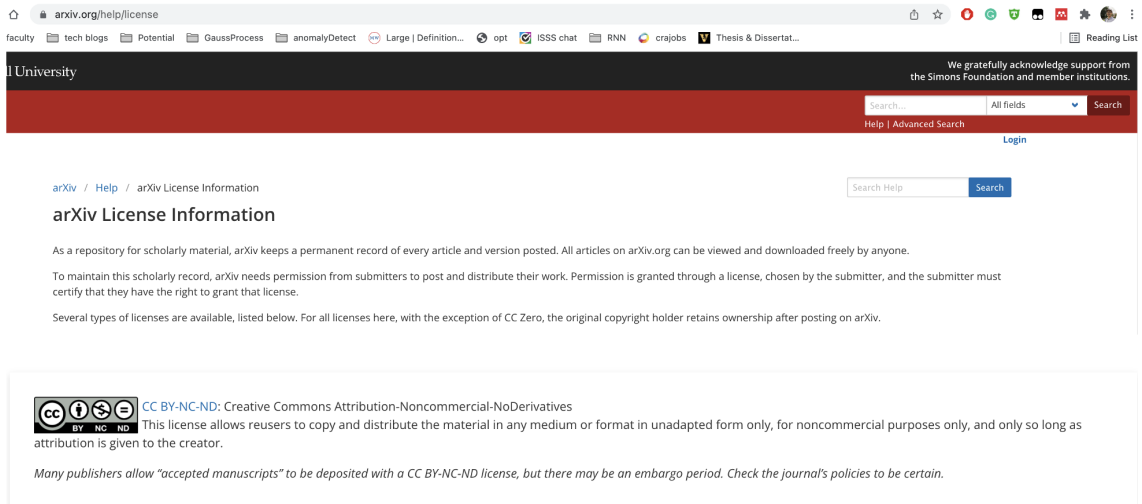
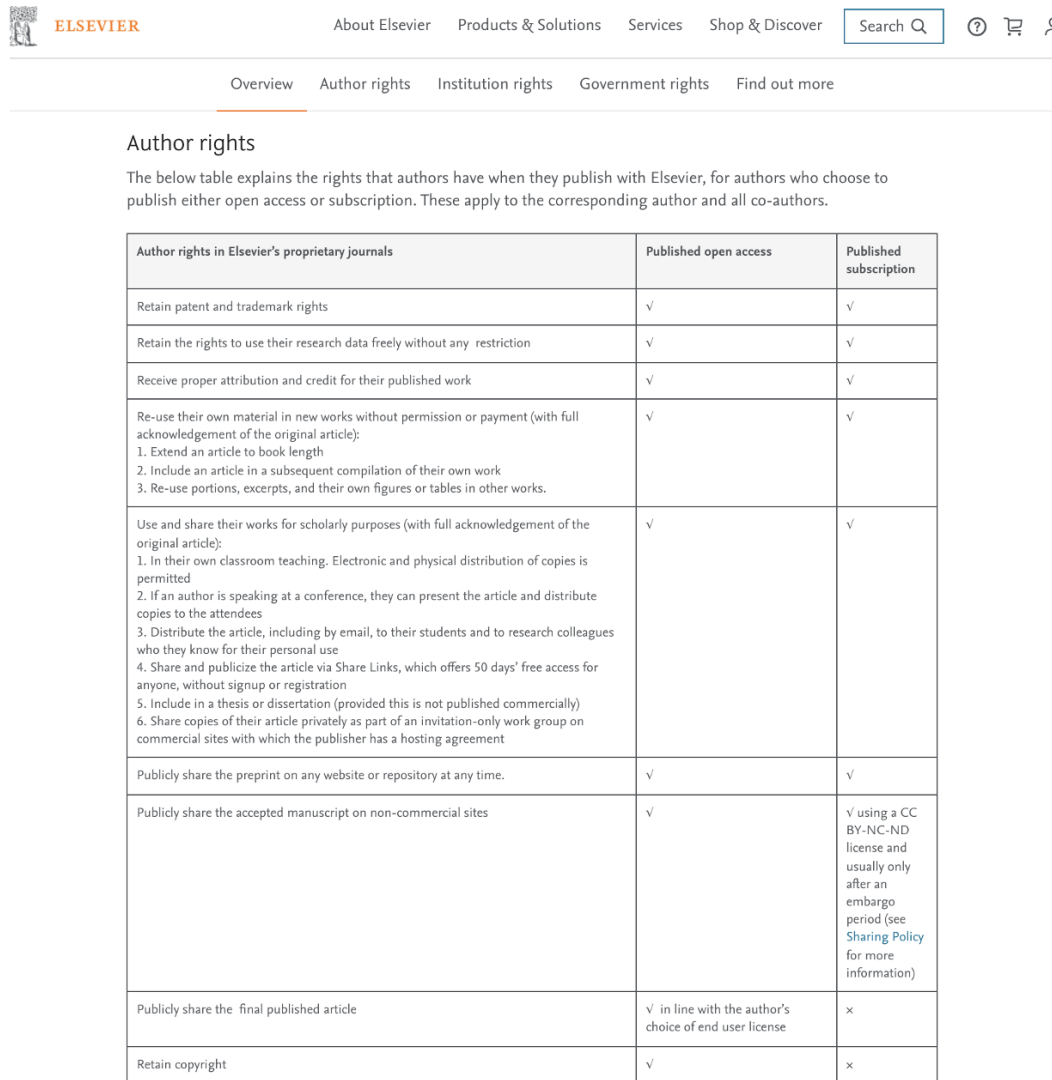


Figure A.1: Copyright from arXiv

A.2 Copyright from Elsevier

Author rights in Elsevier’s proprietary journals (A.2) include re-use portions, excerpts, and their own figures or tables in other works. Our Medical Image Analysis paper (Chapter 5) is under Elsevier publisher.



The screenshot shows the Elsevier website header with navigation links: About Elsevier, Products & Solutions, Services, Shop & Discover, and a search bar. Below the header is a navigation menu with 'Author rights' selected. The main content area is titled 'Author rights' and contains a table explaining the rights authors have when publishing with Elsevier, categorized by 'Published open access' and 'Published subscription'.

Author rights in Elsevier's proprietary journals	Published open access	Published subscription
Retain patent and trademark rights	√	√
Retain the rights to use their research data freely without any restriction	√	√
Receive proper attribution and credit for their published work	√	√
Re-use their own material in new works without permission or payment (with full acknowledgement of the original article): 1. Extend an article to book length 2. Include an article in a subsequent compilation of their own work 3. Re-use portions, excerpts, and their own figures or tables in other works.	√	√
Use and share their works for scholarly purposes (with full acknowledgement of the original article): 1. In their own classroom teaching. Electronic and physical distribution of copies is permitted 2. If an author is speaking at a conference, they can present the article and distribute copies to the attendees 3. Distribute the article, including by email, to their students and to research colleagues who they know for their personal use 4. Share and publicize the article via Share Links, which offers 50 days' free access for anyone, without signup or registration 5. Include in a thesis or dissertation (provided this is not published commercially) 6. Share copies of their article privately as part of an invitation-only work group on commercial sites with which the publisher has a hosting agreement	√	√
Publicly share the preprint on any website or repository at any time.	√	√
Publicly share the accepted manuscript on non-commercial sites	√	√ using a CC BY-NC-ND license and usually only after an embargo period (see Sharing Policy for more information)
Publicly share the final published article	√ in line with the author's choice of end user license	x
Retain copyright	√	x

Figure A.2: Copyright from Elsevier

A.3 Copyright from LNCS

Authors retains the right to use the content for non-commercial internal and educational purposes, etc. Our Chapter 6 is the MICCAI paper under Copyright from LNCS (A.3).

§ 2 Rights Retained by Author

Author retains, in addition to uses permitted by law, the right to communicate the content of the Contribution to other research colleagues, to share the Contribution with them in manuscript form, to perform or present the Contribution or to use the content for non-commercial internal and educational purposes, provided the original source of publication is cited according to the current citation standards in any printed or electronic materials. Author retains the right to republish the Contribution in any collection consisting solely of Author's own works without charge, subject to ensuring that the publication of the Publisher is properly credited and that the relevant copyright notice is repeated verbatim. Author may self-archive an author-created version of his/her Contribution on his/her own website and/or the repository of Author's department or faculty. Author may also deposit this version on his/her funder's or funder's designated repository at the funder's request or as a result of a legal obligation. He/she may not use the Publisher's PDF version, which is posted on the Publisher's platforms, for the purpose of self-archiving or deposit. Furthermore, Author may only post his/her own version, provided acknowledgment is given to the original source of publication and a link is inserted to the published article on the Publisher's website. The link must be provided by inserting the DOI number of the article in the following sentence: "The final authenticated version is available online at [https://doi.org/\[insert DOI\]](https://doi.org/[insert DOI])." The DOI (Digital Object Identifier) can be found at the bottom of the first page of the published paper.

Prior versions of the Contribution published on non-commercial pre-print servers like ArXiv/CoRR and HAL can remain on these servers and/or can be updated with Author's accepted version. The final published version (in pdf or html/xml format) cannot be used for this purpose. Acknowledgment needs to be given to the final publication and a link must be inserted to the published Contribution on the Publisher's website, by inserting the DOI number of the article in the following sentence: "The final authenticated publication is available online at [https://doi.org/\[insert DOI\]](https://doi.org/[insert DOI])".

Author retains the right to use his/her Contribution for his/her further scientific career by including the final published paper in his/her dissertation or doctoral thesis provided acknowledgment is given to the original source of publication. Author also retains the right to use, without having to pay a fee and without having to inform the Publisher, parts of the Contribution (e.g. illustrations) for inclusion in future work. Authors may publish an extended version of their proceedings paper as a journal article provided the following principles are adhered to: a) the extended version includes at least 30% new material, b) the original publication is cited, and c) it includes an explicit statement about the increment (e.g., new results, better description of materials, etc.).

Figure A.3: Copyright from LNCS

A.4 Copyright from SPIE and JMI

This is the permission obtained from SPIE. The email screenshot is shown in A.4. Our content in Chapter 11 is the extension work of SPIE and publication in JMI

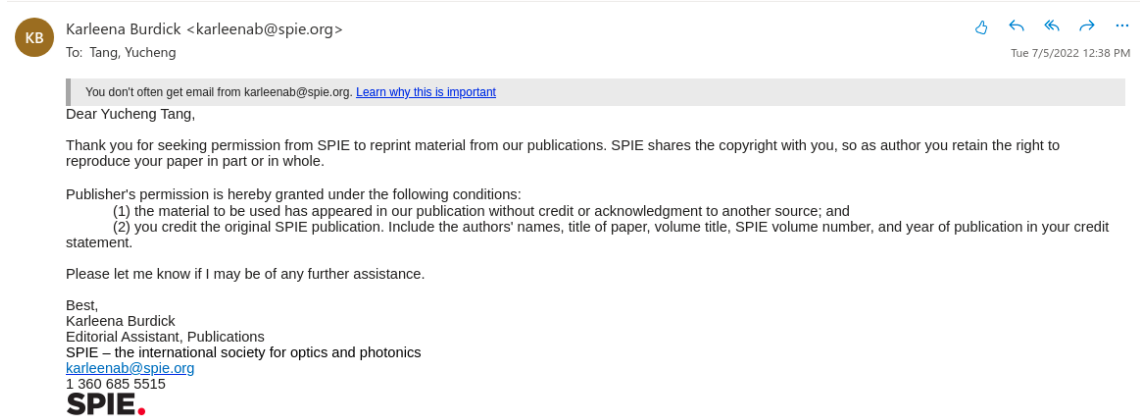


Figure A.4: Copyright from SPIE

A.5 Copyright from IEEE Conference and IEEE TMI

This is the permission obtained from IEEE TMI, Body Part Regression (Chapter 2), 3D transformer (Chapter 7) and self-supervised pre-training (Chapter 8). The email screenshot is shown in A.5.

Copyright

By policy, IEEE owns the copyright to the technical contributions it publishes on behalf of the interests of the IEEE, its authors, and their employers; and to facilitate the appropriate reuse of this material by others. To comply with United States copyright law, authors are required to sign and submit a completed IEEE Copyright Form with their original submissions. IEEE returns to authors and their employers full rights to reuse their material for their own purposes. Electronic submission of the IEEE Copyright Form can be accomplished online. Instructions are available at <http://www.ieee.org/web/publications/rights/copyrightmain.html>.

Figure A.5: Copyright from IEEE

A.6 Copyright from Medical Physics

This is the permission obtained from Medical Physics, Contrast Phase Identification work (Chapter 4). The copyright page screenshot is shown in A.6.

Copying and Other Use

Copyright 2022. American Association of Physicists in Medicine. All rights reserved.

Copying: Single copies of individual articles may be made for private use or research. Authorization is given to copy articles beyond the free use permitted under Sections 107 and 108 of the U.S. Copyright Law, provided the copying fee of \$30.00 per copy per article is paid to the Copyright Clearance Center (CCC), 222 Rosewood Drive, Danvers, MA 01923, USA, www.copyright.com. (Note: The ISSN for this journal is 0094-2405.)

Authorization does not extend to systematic or multiple reproduction, to copying for promotional purposes, to electronic storage or distribution, or to republication in any form. In all such cases, specific written permission from AAPM must be obtained.

Permission for Other Use: Permission is granted to quote from the *Journal* with the customary acknowledgment of the source.

Requests for Permission: Permission for material published in JACMP or Medical Physics Journal may be requested from the article download page.

CHORUS: Through participation in the CHORUS initiative, AAPM will make publicly available the Accepted Manuscript version of an article in response to government or funder requirements 12 months after publication. Unless otherwise noted on the article, the Accepted Manuscript is licensed under the terms of the AAPM Transfer of Copyright Agreement.

Figure A.6: Copyright from Medical Physics