EFFICIENT REPRESENTATION LEARNING FOR OPTICAL IMAGE ANALYSIS

By

Quan Liu

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

Computer Science

August 9, 2024

Nashville, Tennessee

Approved:

Yuankai Huo, Ph.D.

Jason G. Valentine, Ph.D.

Catie Chang, Ph.D.

Jieying Wu, Ph.D.

Sunxing Bao, Ph.D.

Ling Zhang, Ph.D.

ACKNOWLEDGMENTS

My journey as a graduate student with HRLB Lab at Vanderbilt has been both personal and collaborative, filled with gratitude towards those who have supported and inspired me along the way. Firstly, I owe a great debt of thanks to my parents, Lihua Miao and Lijun Liu, whose unwavering support, love, and encouragement have guided me through the world of science and research. Their belief in me and their encouragement to explore the wonders of science and pursue my dreams have been invaluable.

I am immensely grateful to Yuankai Huo, my mentor, whose guidance has played a pivotal role in shaping me into the researcher I am today. His mentorship has not only empowered me to think independently but has also fostered a strong foundation for my research endeavors. It was Yuankai who encouraged me to apply to Vanderbilt and join the HRLB Group, a decision for which I am deeply thankful. Under his guidance, I have explored the fields of medical image analysis and computer vision, discovering profound satisfaction in applying machine learning to healthcare.

I also extend my gratitude to Dr. Valentine for our collaboration on the meta-optic project and for our joint efforts in publishing our work in a top journal. Working with Dr. Valentine has been both enriching and rewarding, and I am grateful for the opportunity to contribute to such impactful research.

I am also thankful for the warm and supportive environment fostered by the HRLB Lab at Vanderbilt. Ruining, your leadership and expertise have been invaluable, and I am grateful to have had the opportunity to work alongside you. Tianyuan, your assistance with both research and life challenges has been invaluable, and I appreciate your unwavering support. Cathy, your passion for learning and curiosity has inspired me to approach research with a thorough and inquisitive mindset.

Lastly, I consider myself fortunate to be part of the global community of scientists in the field of medical image analysis. Our collective efforts have the potential to significantly impact human health and contribute to a better world, one small milestone at a time.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS . . . . **iii**

LIST OF TABLES . . . . **x**

LIST OF FIGURES . . . . **xi**

1 Introduction **1**

1.1 Overview 1
1.2 Challenges on Optical Image Analysis 3
1.2.1 Challenge on Label Efficient Semantic Segmentation on Microscopic Image Analysis 3
1.2.2 Challenge on Efficient Feature Representation Learning for Medical Optical Imaging 4
1.2.3 Challenge on Energy Efficient Representation Learning for Meta-optics . 6
1.3 Label Efficient Representation Learning on Biological Image 6
1.3.1 Unsupervised Semantic Segmentation in Microscopy Imaging 6
1.3.2 Image Synthesis without Annotations 7
1.3.3 Microscope Image Segmentation and Tracking 7
1.4 Efficient Feature Representation Learning Methods in Medical Optical Imaging . 8
1.4.1 Unsupervised Representation Learning Model by Contrastive Learning . 8
1.4.2 Efficient Model Training Adaption on Limited GPU Resources 9
1.4.3 Efficient Fine-tuning Methods for Pretrained Model on Downstream Task 9
1.4.4 Efficient Multi-Modality Representation Learning on Medical Image . . . 10
1.5 Energy Efficiency Representation Learning for Meta-optic 10

2 Label Efficient Representation Learning for Biological Optical Imaging **12**

2.1 GAN based Unsupervised Segmentation on Microscopic Image 12
2.1.1 Introduction 12
2.1.2 Methods 13
2.1.2.1 Cycle-consistent image synthesis 13
2.1.2.2 Micro- and macro-level matching 15
2.1.2.3 Image segmentation 16
2.1.3 Data and experimental design 16
2.1.3.1 Microvilli images 16
2.1.3.2 Experimental design 17
2.1.4 Results 20
2.1.5 Conclusion 20
2.2 Annotation-free Synthetic Instance Segmentation and Tracking for Microscope Video 21

2.2.1 Introduction 21
2.2.2 Related Work 24
2.2.2.1 Image synthesis 24
2.2.2.2 Microscope image segmentation and tracking 24
2.2.3 Methods 25
2.2.3.1 Unsupervised image-annotation synthesis 25
2.2.3.2 Video synthesis 26
2.2.3.2.1 Microvilli simulation 27
2.2.3.2.2 HeLa cell simulation 29
2.2.3.3 Annotation refinement for HeLa cell video simulation 29
2.2.3.3.1 Binary mask generation 30
2.2.3.3.2 Annotation deformation (AD) 30
2.2.3.3.3 Annotation cleaning (AC) 30
2.2.3.4 Instance segmentation and tracking 31
2.2.4 Experimental design 31
2.2.4.1 Instance segmentation and tracking on microvilli video 31
2.2.4.1.1 Data 31
2.2.4.1.2 Experimental design 31
2.2.4.2 Instance segmentation and tracking on HeLa cell video 32
2.2.4.2.1 Data 32
2.2.4.2.2 Experimental design 32
2.2.4.3 Evaluation matrix 33
2.2.5 Results 33
2.2.5.1 Instance segmentation and tracking on microvilli videos 33
2.2.5.2 Instance segmentation and tracking on HeLa cell videos 35
2.2.6 Discussion 35
2.2.7 Conclusion 37

3 Efficient Feature Representation Learning for Medical Optical Imaging 38

3.1 Simple Triplet Representation Learning on Histopathology Image with a Single GPU 38
3.1.1 Introduction 38
3.1.2 Methods 40
3.1.2.1 Multi-view nature of medical images 41
3.1.2.2 Triplet representation learning 41
3.1.2.3 Expand batch size via mix precision training 42
3.1.3 Data and Experiments 43
3.1.3.1 Data 43
3.1.3.2 Supervised learning 43
3.1.3.3 Training contrastive learning benchmarks 44
3.1.3.4 Training the proposed SimTriplet 44
3.1.3.5 Linear evaluation (fine tuning) 45
3.1.4 Results 46
3.1.4.1 Model classification performance. 46
3.1.4.2 Model performance on partial training data. 46
3.1.5 Conclusion 47
3.2 Integrate Memory Efficiency Methods for Self-supervised Learning on Pathological Image Analysis 48

3.2.1 Introduction 48
3.2.2 Method 50
3.2.2.1 BYOL 50
3.2.2.2 Data parallel computing 51
3.2.2.3 Mixed precision training 51
3.2.2.4 ActNN 51
3.2.2.5 In-place Operation 51
3.2.3 Data and experimental design 52
3.2.3.1 Data 52
3.2.3.1.1 Annotated data 52
3.2.3.1.2 Unlabeled data 52
3.2.3.2 Experimental design 52
3.2.3.2.1 Oracle 52
3.2.3.2.2 In-place 53
3.2.3.2.3 Mixed precision training + In-place 53
3.2.3.2.4 ActNN + In-place 53
3.2.3.2.5 Linear evaluation (fine-tuning) 53
3.2.4 Results 53
3.2.4.1 Model classification performance 54
3.2.4.2 Model training speed 54
3.2.5 Discussion 55
3.3 Leverage Supervised and Self-supervised Pretrain Models for Pathological Survival Analysis via a Simple and Low-cost Joint Representation Tuning 56
3.3.1 Introduction 56
3.3.2 Methods 58
3.3.2.1 Supervised and self-supervised pretraining 58
3.3.2.2 Joint representation tuning 58
3.3.2.3 Evaluate different strategies of using pretrained model 58
3.3.3 Experiments 59
3.3.3.1 Data description 59
3.3.3.1.1 Survival prediction task 59
3.3.3.1.2 Classification task 59
3.3.3.2 Experimental setting 59
3.3.4 Results 60
3.3.4.1 Performance on classification and survival prediction 60
3.3.4.2 Computational resource 61
3.3.5 Ablation Studies 62
3.3.5.1 Different strategies of using pretrained models 62
3.3.5.2 Effect of transformer 62
3.3.5.3 Effect of data augmentation 63
3.3.6 Conclusion 63
3.4 Bayesian-based Multimodal Multi-level Fusion on Colorectal Cancer Microsatellite Instability Prediction 65
3.4.1 Introduction 65
3.4.2 Method 66
3.4.2.1 Bayesian-based multi-modality fusion model 67
3.4.2.2 MSI prediction on single modality 68
3.4.2.3 Model prediction fusion on multiple levels 69

3.4.2.3.1 Decision level multimodal fusion 69
3.4.2.3.2 Feature level multi-modal fusion 70
3.4.3 Experiments 71
3.4.3.1 Dataset 71
3.4.3.2 Experimental Design 71
3.4.4 Results 73
3.4.5 Conclusion 73
3.5 Multi-Level Text-Guided Representation End-to-End Learning for Whole Slide Image Analysis 75
3.5.1 Introduction 75
3.5.2 Related Work 77
3.5.2.1 Multi-instance learning 77
3.5.2.2 Attention sampling 78
3.5.2.3 Visual language model in WSI analysis 78
3.5.3 Methods 79
3.5.3.1 Global alignment 80
3.5.3.2 Local alignment 81
3.5.3.3 End-to-end training with image sampler 82
3.5.4 Experiments 82
3.5.4.1 Data description 82
3.5.4.2 Data preprocessing 83
3.5.4.3 Network architectures 83
3.5.4.4 Training details 84
3.5.4.5 Baseline experiments for comparison 85
3.5.5 Results 86
3.5.5.1 KIRC 86
3.5.5.2 GBMLGG 86
3.5.5.3 Evaluation for multi-level text alignments 87
3.5.5.4 Visualization 88
3.5.6 Conclusion 88

4 Energy Efficient Representation Learning for Meta-optics 89

4.1 Digital Modeling on Large Kernel Metamaterial Neural Network 89
4.1.1 Introduction 89
4.1.2 Related work 91
4.1.2.1 Models with large kernel convolution 91
4.1.2.2 Model compression and re-parameterization 93
4.1.2.3 Optical neural network 94
4.1.3 Method 94
4.1.3.1 Large kernel re-parameterization 94
4.1.3.2 Meta-optic adaptation 97
4.1.3.3 Model compression 99
4.1.4 Data and experimental design 100
4.1.4.1 Data description 100
4.1.4.2 Large kernel re-parameterization 101
4.1.4.3 Meta-optic model adaptation 101
4.1.4.4 Model compression efficiency 102

4.1.5 Result 102
4.1.5.1 Large re-parameterized convolution performance 102
4.1.5.2 Performance of model adaptation 104
4.1.5.3 Ablation studies 105
4.1.5.4 LMNN efficiency and speed evaluation 105
4.1.6 Discussion 106
4.1.7 Conclusion 107
4.2 High-speed lightweight image segmentation by remodeling multi-channel meta-imagers 108
4.2.1 Introduction 108
4.2.2 Related work 110
4.2.2.1 Large kernel convolution design 110
4.2.2.2 Optic neural network 111
4.2.2.3 Convolution neural network model compression 111
4.2.3 Method 112
4.2.3.1 Large convolution design with multiple path design 112
4.2.3.2 Model compression with sparse convolution 113
4.2.4 Data and experimental design 115
4.2.4.1 Data description 115
4.2.4.2 Data generation with foundation model 115
4.2.4.3 Large kernel digital design on segmentation model 115
4.2.4.4 Model design with optic constrain 116
4.2.4.5 Model compression efficiency 116
4.2.5 Result 116
4.2.5.1 Segmentation performance on portrait dataset 116
4.2.5.2 Segmentation performance on car dataset 117
4.2.5.3 Ablation studies 118
4.2.5.4 Model compression 119
4.2.6 Discussion 120
4.2.7 Conclusion 121

5 Contribution and Future Work 122

5.1 Contributions 122
5.1.1 Contribution on annotation-free semantic segmentation with microscopic image 122
5.1.2 Contribution on annotation-free synthetic instance segmentation and tracking for microscope video 122
5.1.3 Contribution on simple triplet representation learning (SimTriplet) approach on pathological images 123
5.1.4 Contribution on memory efficiency methods for self-supervised learning on pathological image analysis 123
5.1.5 Contribution on leverage the trained supervised and self-supervised models for pathological image survival analysis 123
5.1.6 Contribution on multimodal multi-level fusion on Colorectal cancer microsatellite instability prediction 124
5.1.7 Contribution on multi-level text-guided representation end-to-end learning for whole slide image analysis 124

5.1.8 Contribution on large convolution kernel design on meta-optics image classification 125
5.1.9 Contribution on large convolution kernel design on meta-optics image segmentation 125
5.2 Future work 125
5.2.1 Exploring multi-modalities assisted representation learning on pathology image 125
5.2.2 Exploring the foundation model ability on medical image analysis 126
5.2.3 Exploring the representation learning on super-resolution image 126
5.2.4 Exploring the Large Vision Model (LVM) on medical image analysis . . . 127

References 128

A Copyright from Publishers 149

A.1 Copyright from SPIE and JMI 149
A.2 Copyright from Elsevier 150
A.3 Copyright from LNCS 151
A.4 Copyright from IS&T 152

LIST OF TABLES

Table Page

2.1 The average Dice values of different experiments. 19
2.2 DET, SET and TRA values of different experiments on microvilli video. 34
2.3 DET, SET and TRA values of different experiments on HeLa cell video. 36

3.1 Classification performance. 47
3.2 Balanced Acc of using different percentage of annotated data. 47
3.3 Classification performance. 54
3.4 Survival prediction performance on TCGA-GBMLGG ROI dataset. 62
3.5 Classification performance on NCT-CRC-HE dataset. 62
3.6 Comparison of volumetric analysis metrics between the proposed method and the state-of-the-art clinical study on kidney components. 63
3.7 AUC on MSI prediction 73
3.8 Ablation study for pathology feature aggregation and feature-level fusion strategy 74
3.9 Cancer grade classification and survival prediction results on KIRC dataset. Acc represents the accuracy of grade classification, while the C-Index evaluates the survival prediction performance. 83
3.10 Cancer grade classification and survival prediction results on GBMLGG dataset. 'Acc' represents the accuracy of grade classification, while the C-Index evaluates the survival prediction performance. 84
3.11 Ablation study for multi-level text alignments is shown in this table. The accuracy of grade classification (ACC) and the survival prediction performance (C-Index) are presented. 88

4.1 Large re-parameterized convolution experiment results 103
4.2 Metasurface fabrication 105
4.3 Segmentation performance on EG1800 117
4.4 Segmentation performance on EG1800 after model compression 117
4.5 Segmentation performance on car dataset 118
4.6 Segmentation performance on car dataset after model compression 118

LIST OF FIGURES

Figure Page

2.1 This figure shows the general framework of the image synthesis. CycleGAN is used to perform synthesis between real images and simulated masks. In micro-level matching, the green protein channel in fluorescence images is used to achieve the cell counting automatically, which provides the rough numbers of microvilli in the real images. Then, the corresponding simulated masks with the same number of sticks are generated from the simulator when forming a mini-batch for training. In macro-level matching, the numbers of sticks in the simulated masks are randomly generated from the prior global distribution, without mini-batch level correspondence. 14

2.2 This figure demonstrates four experimental designs of providing different augmentations of masks for training the CycleGAN. The first design employs the binary mask directly, while the remaining three designs utilize different augmentation strategies. 15

2.3 This figure shows the segmentation pipeline. The input images of the U-Net model are the synthetic fake images from trained CycleGAN's generator, while the annotations are the simulated masks. Note that, no manual annotations are used in our training either for CycleGAN or U-Net, as an unsupervised framework. 16

2.4 The synthesis results of different experimental designs are provided in this figure. The first column is the initial simulated masks with different numbers of objects (sticks). The middle columns exhibit the synthetic images from the masks using different augmentation strategies. The last column is five randomly selected real images, which are unpaired to the masks in CycleGAN framework. 18

2.5 The final segmentation results from U-Net are presented in this figure. Each row shows the segmentation results at different frames in a microvilli video. 19

2.6 The upper panel shows the existing pixel-embedding deep learning based single-stage instance segmentation and tracking method, which is trained by real microscope video and manual annotations. The lower panel presents our pro-posed annotation-free ASIST method, with synthesized data and annotations from adversarial simulations. 22

2.7 Real and synthetic video of Hela cell and microvilli consisting of three aspects: shape, appearance and dynamics. The ”shape” is defined as the underlying shape of the manual annotations. The ”appearance” is defined by the various appearances of objects. The ”dynamics” indicates the mitigation of cellular and subcellular objects. 23

2.8 This figure shows the proposed ASIST method. First, CycleGAN based image-annotation synthesis is trained using real microscope images and simulated annotations. Second, synthesized microscope videos are generated from simulated annotation videos. Last, an embedding based instance segmentation and tracking algorithm is trained using synthetic training data. For HeLa cell videos, a new annotation refinement step is introduced to capture the larger shape variations. . 26

2.9 The left panel shows real microscope videos as well as manual annotations. The right panel presents our synthetic videos and simulated annotations. 27

2.10 The upper panel shows the CycleGAN that is trained by real images and simulated annotations with Gaussian blurring. The lower panel shows the CycleGAN that is trained by the same data without Gaussian blurring. The Generator B is used to generate synthetic videos with larger shape variations from circle representations, while the Generator A* generate sharp segmentation for the annotation registrations. 28

2.11 This figures shows the workflow of the annotation refinement approach. The simulated circle annotations are fed into Generator B to synthesize cell images. We used Generator A* in Fig.2.10 to generate sharp binary masks from synthetic images. Then, we registered simulated circle annotations to binary masks to match the shape of cells in synthetic images. Last, an annotation cleaning step was introduced to delete the inconsistent annotations between deformed instance object masks and binary masks. 28

2.12 This figure shows the instance segmentation and tracking results of the real testing microvilli video. 34

2.13 This figure shows the instance segmentation and tracking results on the real HeLa cell testing video. 35

3.1 Comparison of contrastive learning strategies. The upper panel compares the proposed SimTriplet with current representative contrastive learning strategies. The lower panel compares different approaches via a table. 39

3.2 Network structure of the proposed SimTriplet. Adjacent image pairs are sampled from unlabeled pathological images (left panel) for triplet representation learning (right panel). The GigaPixel pathological images provide large-scale ”positive pairs” from nearby image patches for SimTriplet. Each triplet consists of two augmentation views from m_1 and one augmentation view from m_2. The final loss maximizes both inter-sample and intra-sample similarity as a representation learning. 40

3.3 Compare SimTriplet with SimSiam. SimSiam network maximizes intra-sample similarity by minimizing the distance between two augmentation views from the same image. The proposed SimTriplet model further enforces the inter-sample similarity from positive sample pairs. 41

3.4 Visualization of classification results. One tissue sample is manually segmented by our dermatologist (via QuPath software) to visually compare the classification results. The contrasting learning achieved superior performance compared with supervised learning, even using only 1% of all available labeled data. 45

3.5 t-SNE plot of abstracted feature by SimTriplet model. The abstracted feature is shown in t-SNE plot. Different color dots represent different tissue types. 46

3.6 GPU maximizing efficiency comparison. Blue circles use no GPU memory-efficient methods. ”bs” means the batch size used in model training. MP is mixed precision training. Method on the lower right is preferred which achieves a larger batch size utilizing less GPU memory. 49

3.7 GPU maximizing efficiency methods pipelines. (a) Mixed precision training: In the forward and backward pass, FP16 is used for computing which halves bandwidth and GPU occupancy rate. (b) ActNN: Compress activations before storage and decompress activation for backpropagation. (c) Data parallel computing: Multiple GPU devices process data concurrently and update the shared model weights. 50

3.8 GPU maximizing efficiency methods training speed chart. ”bs” means training batch size used in model training. Oracle is implemented without GPU memory-efficient strategy and run parallel on multiple GPU devices. 55

3.9 Model pipeline. In the top section, self-supervised model is pretrained with pathology WSIs and finetuned on pathology images for survival prediction. Middle section is supervised pretrained model with natural images of ImageNet and finetuned on pathology images. Our JRT method aggregates both pretrained models to achieve better downstream task performance. 57

3.10 Finetuning strategies on pre-trained backbone. This figure shows the three prevalent finetuning strategies: (1) No-freeze, (2) Encoder-freeze, and (3) Feature-direct using ResNet and TranPath. The highlighted method is the proposed JRT strategy. 60

3.11 Visual comparison of different strategies. This figure shows the C-Index versus the model training time required for an epoch. The size of the blobs is proportional to the number of GPU memory consumption. 61

3.12 Our proposed M^2Fusion model. Multimodal data, WSI, and CT images are preprocessed to pathology image patches and CT tumor ROI, respectively. Embeddings are extracted by encoder E_p and E_r. $*$ means the model is well-trained and frozen in pipeline training. $\mathscr{P}_P$ is the pathology uni-model performance $\mathscr{P}(P_{ath})$. P_R is the radiology uni-model performance $\mathscr{P}(R_{ad})$. $\mathscr{P}_F$ is the feature level fusion model probability distribution under pathology and radiology guidance $\mathscr{P}(F_{ea}|P_{ath}R_{ad})$. The final fusion model by P_P, P_R and P_F is $\mathscr{P}(F_{ea}P_{ath}R_{ad})$ in Eq.3.8 66

3.13 Baseline experiments on multimodal fusion. A. Decision level multimodal fusion, $\mathscr{P}(P_{ath}R_{ad})$ in Eq.3.5. B. Radiology-guided feature-level fusion, probability distribution follows $\mathscr{P}(F_{ea}|R_{ad})$. ’*’ means the model is well-trained and frozen in pipeline training. 69

3.14 Data visualization of the dataset. First row shows two modalities image from MSS subjects. The second row shows data from MSI subject. 70

3.15 Comparison between multi-instance learning, pathologist diagnosis, and our proposed mTREE. (a) Traditional multi-instance learning needs to process all patches without patch selection. (b) Pathologists in the diagnosis process focus on the most essential patches selected by manual efforts. (c) Our proposed mTREE generates text-guided attention to sample efficiently without manual annotation. 76

3.16 This figure demonstrates the proposed mTREE pipeline. The upper panel shows the text process flow. The text encoder is frozen with pre-trained weights. Text feature T_0 is used for global alignment and alignment of image patch features. The lower panel shows the WSI analytic flow. The attention model learns an attention map from the WSI in low resolution. The attention map aligns with the text feature T_0. The image patches tiled up from high-resolution WSI are ranked by attention score. The image features $I_0, I_1 ... I_k$ abstracted from image patches with higher attention scores are aggregated with text feature T_0. 79

3.17 This figure presents the principle of multi-level text guidance. Global-level text guidance (upper panel) aligns the attention map from images and text. Image attention map is learned from low-resolution WSI, while text attention is projected from the text feature T_0. Local-level text guidance (lower panel) performs patch selection by computing the cosine similarity distance to the text feature T_0 and aggregates features from both image and text. 80

3.18 This figure presents the visualization of WSI-level attention and the automatically derived diagnosis patches. For WSIs in the TCGA-KIRC dataset and TCGA-GBMLGG dataset, the attention map (middle panels) is learned from WSI (left panels), highlighting essential tissue regions. Essential image patches (right panels) are selected according to the attention score. The image boundary color indicates the according attention score. 87

4.1 This study provides a digital modeling platform for designing and optimizing a metamaterial neural network (MNN). The proposed large kernel metamaterial neural network (LMNN) is able to maximize the performance of an MNN without introducing extra computational complexity during the inference stage. . . . 91

4.2 The upper panel (a) shows the conventional CNN model on the image classification task with Batch Normalization (BN) and Multilayer Perceptron (MLP). The lower panels (b) present our proposed LMNN method with digital design and optic implementation with depthwise convolution (DWC) layer. The large kernel re-parameterization efficiently achieves a large receptive field with a multi-branch multi-layer structure. Physical constraints are modeled via the meta-optic adaptation. The multi-branch multi-layer model is further compressed to a single-layer LMNN. (c) The digital design is fabricated as a real meta-optic device for inference. The red arrow shows the main pipeline to build the LMNN. The green arrow shows the image processing path in meta-optic imaging system. 92

4.3 An overview of our proposed large convolution kernel block with re-parameterization is presented. Learnable scaling factors are employed to mimic the scaling function of batch normalization. In the inference stage, the block can be converted to a single convolution layer. 'FC' refers to fully connected layer in the figure. . . 95

4.4 Adaptation for meta-optic implementation. (a) To implement the kernel with negative weight, we split the kernel into the positive kernel and negative kernel and subtract from their feature map. (b) The non-linear layer needs to be removed from the parallel convolution path. 99

4.5 The meta-optic devices simulation and implementation platform. (a) Optic system for meta-optic lens test. The components in the figure are: Light source: Tungsten Lamp; filter: Wavelength filter; Pol: Polarizer; SLM: Spatial light modulator; Condenser: Lens to focus light on the SLM; MSs: Metasurfaces; Ob: Objective lens. (b) Measured meta-optic kernel weight point spread function, used for optical convolution with the imaged object. (c) Theoretical meta-optic kernel weight point spread function by simulation. 100

4.6 Performance of the Large Convolution WideResNet-101 model on the STL-10 dataset. The model was assessed using varying convolution kernel sizes, ranging from 7×7 to 31×31. The LMNN consistently outperforms, highlighting the benefits of utilizing larger convolution kernels. 104

4.7 (a) The large re-parameterized convolution model performance with different layer numbers and channel numbers. (b) Large convolution kernel efficiency evaluation. The circle in different colors shows different convolution layer structures. The shadow area is the model structure that can be fabricated. The circle area shows the FLOPs ratio of the layer implemented by meta-optic material. x-axis is the model FLOPs except the layer to be fabricated. (c)Model inference time between the baseline digital model and hybrid model. The orange bar in the figure shows the time used our model. 104

4.8 Plot of ablation study on LMNN. (a) Evaluating non-negative weight effect on model performance. (b) Measuring the effect of model bandwidth and weight precision effect on model prediction accuracy. ’Pr’ in Figure (b) means precision. 105
4.9 This study provides a hybrid pipeline for designing and optimizing a large kernel metamaterial neural network (MNN). The proposed Meta-imager is efficient for segmentation tasks with fewer FLOPs in computation. 109
4.10 Lightweight segmentation model with hybrid meta optics design. 113
4.11 Model compression on segmentation model digital processing part. 114
4.12 Model ablation study. Left panel: trade-off between input image size and channel number of convolution layer. Right panel: model efficiency visualization comparing model FLOPs and mIoU. 119
4.13 Model compression performance. Left panel: origin model, meta-imager, and compressed model parameters comparison; right panel: model performance after compression. 119

A.1 Copyright from SPIE 149
A.2 Copyright from Elsevier 150
A.3 Copyright from LNCS 151
A.4 Copyright from IS&T 152

CHAPTER 1

Introduction

1.1 Overview

Optical image analysis is a fundamental task in the field of computer vision for extracting information from images. It involves various types of optical images, including microscopic images, pathology images, and meta-optic images, that are commonly utilized in computer vision tasks. Microscopic images, captured using a microscope, provide magnified views of the micro-world and have proven useful in understanding biological scenes at a cellular or subcellular level (Wang et al., 2022b). These images are crucial in studying micro-environments of the human body, such as leukocytes (Saraswat and Arya, 2014) or hippocampus brain cell (Hore et al., 2015). Histopathology images represent another critical type of optical image, which serves as primary data for cancer diagnosis in clinical practice. These images can provide information about the molecular details of the tissue being examined (Lee et al., 2021). To maintain the microenvironment details, histopathology images are generated by digital slide scanners, which create high-resolution images that require significant computing resources. Finally, meta-optic images are generated by metasurface systems that manipulate light and polarization to create images. These images resemble natural images and can perform mathematical operations with appropriate metasurface design. Since metasurfaces can be manually designed for image processing, meta-optic images offer opportunities for conducting specific types of mathematical operations, making them particularly valuable in computer vision research.

The utilization of optical images in representation learning poses significant challenges when compared to well-established natural image datasets such as ImageNet and CIFAR10. Natural image datasets are known to have well-organized annotations and sufficient data per category, which facilitates efficient representation learning and optimal model performance. In contrast, optical images, such as those obtained from microscopic images, pathology images, and meta-optic images from meta-material, are characterized by unique complexities that make them unsuitable for traditional representation learning approaches. The high requirements for representation learning on optical images make the tasks prohibitively time-consuming and expensive, hindering efficient and effective learning. As a result, specialized techniques such as data augmentation, transfer learning, and domain adaptation are necessary to address these challenges and achieve satisfactory model performance when dealing with optical images.

In medical image research, microscopic images generated from microscopes are revolutionizing the field of biological diagnostics and pathology research (Wang et al., 2022b). While these images provide valuable insights into the micro-world, they also present unique challenges for representation learning in computer vision (Weinstein, 2018). Unlike natural image datasets such as ImageNet (Krizhevsky et al., 2017), which have well-organized annotations and sufficient data per category (Sun et al., 2017), microscopic images have limited numbers of similar organisms and cells, making it challenging to obtain sufficient annotated data for representation learning. Furthermore, the large image size and dataset size of giga-pixel level images in pathology research pose further challenges for annotation and model training (Lu et al., 2021b), contributing to long training times and low task performance (Marini et al., 2021). As such, specialized representation learning strategies are necessary to achieve satisfactory model performance in optical image analysis.

In recent years, the availability of large amounts of patient data has led to significant benefits in representation learning for computer-aided diagnosis in the field of pathology. With the growing public access to pathology image datasets, such as TCGA and PAIP, there has been an increased potential for training machine learning models. The TCGA dataset, for example, contains over 20,000 normal and cancer samples, providing sufficient data for model representation training. However, histopathology data analysis presents a unique challenge due to the large volume of samples, with each histopathology image at a gigapixel level. To overcome this issue, it is not feasible to feed entire histopathology images to the model (Madabhushi, 2009). Instead, cropping histopathology images into patches of normal size (e.g., 256×256) generates tens of thousands of patches for each image. However, using currently available deep learning models and computing resources, training a representation model using these patches would take weeks and is unscalable and inefficient in both academia and industry.

To effectively analyze large-scale image data, it is necessary to extract hidden representations from meta-optics images, which are generated by optical processors and captured by cameras. While traditional image representation models are well-suited for medical images such as microscopic or histopathology images, they encounter efficiency issues when applied to rapidly developing nature image datasets such as ImageNet. This is due to the significant amount of time and energy required for training traditional image representation models, which can be limiting when computing resources are insufficient. Meta-optics images provide a solution to this problem by reducing the energy cost and complexity of mathematical operations in digital neural networks. These mathematical operations, also known as floating-point operations (FLOPs), are a significant bottleneck in popular DNNs (Chen et al., 2017), (Neshatpour et al., 2019). The optical system can be implemented using either free-space (del Hougne et al., 2020; Mennel et al., 2020; Hamerly et al., 2019) or chip-based (Zhang et al., 2021), (Wu et al., 2021) approaches, both of which have achieved low power consumption and ultra-fast speed with meta-optics.

In computer vision tasks such as classification (Krishna et al., 2018), image segmentation (Minaee et al., 2022), and object detection (Gaszczak et al., 2011), the performance of the model heavily depends on the quality of the abstracted representation (Bengio et al., 2013). The abstracted image representation entangles the hidden differences and similarities between images. Achieving promising performance on computer vision tasks using representation learning typically requires large-scale task-specific data, sophisticated model structures, or large computing resources. For instance, ImageNet (Deng et al., 2010), the most widely used large-scale image database, contains 3.2 million images, providing sufficient data to support complex model structures and robust performance on vision tasks. In addition to data, the performance of optical image analysis is strongly related to the model structure and volume (He et al., 2016). Another critical aspect of representation learning for optical image analysis is the computing resources used for model training. Training complex models on large-scale datasets requires massive computing resources, such as GPUs, to enable quick model updates and convergence.

This thesis focuses on addressing the challenges of efficient representation learning in three categories of optical images: microscopic images, pathology images, and meta-optic images, using new methods.

1.2 Challenges on Optical Image Analysis

1.2.1 Challenge on Label Efficient Semantic Segmentation on Microscopic Image Analysis

Semantic segmentation is an important task for microscopic image analysis. Semantic segmentation aims to segment the target object apart from the background context. Multiple traditional methods have been established to handle the semantic segmentation task. Most of them are based on image intensity like graph-cut (Pauchard et al., 2016), Gaussian Mixture model (GMM) (Ban et al., 2018), and watershed (Nguyen et al., 2003). Due to the intensity-based methods needing pixel-wised annotation, the traditional methods are not scalable for large-scale microscopic image datasets. As the microscopic image increases, the deep learning model plays an important role in dealing with large amounts of data and improving thmod'l'sel performance. Due to the semantic segmentation on microscopic images that needs annotation, the limitation makes model training resource intensive. As an unsupervised deep learning method, CycleGAN is a breakthrough framework for the generative adversarial network (GAN). (Dunn et al., 2019; Gadermayr et al., 2019; Ihle et al., 2019b) proposed to use CycleGAN on cellular segmentation. For the subcellular organism segmentation, the CycleGAN model is limited by the objects overlapping and the sub-cellular dynamic nature.

Besides the microscopic image segmentation, the quantitative analysis on microscopic videos requires instance segmentation and tracking tasks on cellular and sub-cellular objects. Compared with segmentation on microscopic images, instance tracking on microscopic videos needs more decent annotation. Traditional object tracking tasks normally consisted of two parts: (1) target object segmentation from the background on each frame, and (2) associating objects in different frames. To simplify the two-stage method to a single step, the pixel-embedding method (Zhao et al., 2021) provides the solution by minimizing the representation distance from the same objects in different frames and maximizing the representation distance from different objects. The method is resource-sensitive, which not only needs object annotation on each frame but also requires consistent annotation across all frames. Considering the high-density dynamic objects in microscopic images, the annotation is more resource-intensive. The representation learning on microscopic images is limited by resource-intensive object annotations.

1.2.2 Challenge on Efficient Feature Representation Learning for Medical Optical Imaging

Over the past years, the dramatic data increases of medical optical images and developments of computation algorithms offer the deep learning model abilities to accelerate diagnosis and guide treatment. Medical images, like histopathology images or CT images, need informatics feature representations for clinically relevant tasks. Feature extraction from medical images is an essential step for computer-assisted methods. Based on the feature extraction algorithms and models in nature image datasets, the performance is consistently promising for pathology images (Hoffer and Ailon, 2015). However, the supervised training strategy on medical image feature extraction needs massive annotation and computing resources. The size of pathology images is at the gigapixel level, which makes the feature extraction even more resource-intensive.

The representation learning model is resource-extensive when training with large-scale datasets. For medical images like Whole Slide Images (WSIs), it is not adaptable to feed large-scale data into the model directly. To minimize the need to process WSI directly, a well-accepted learning strategy is to first learn local image features through unsupervised feature learning, and then aggregate the features with multi-instance learning or supervised learning (Hoffer and Ailon, 2015). On the other hand, models that achieve state-of-the-art performance require sufficient computing resources. For example, within contrastive learning, the primary limitation is that contrastive learning methods need a large batch size to learn the similarity and dissimilarity between samples within the same batch. Limited computing resources are a barrier.

The large-scale pre-trained models from terabyte-level (TB) data are now broadly used in feature extraction, model initialization, and transfer learning in pathological image analyses. Most existing studies have focused on developing more powerful pre-trained models, which are increasingly unscalable for academic institutes. Very few, if any, studies have investigated how to take advantage of existing, yet heterogeneous, pre-trained models for downstream tasks. As an example, our experiments elucidated those self-supervised models (e.g., contrastive learning on the entire The Cancer Genome Atlas (TCGA) dataset) achieved superior performance compared with supervised models (e.g., ImageNet pre-trained) on a classification cohort. Surprisingly, it yielded an inferior performance when it was translated to a cancer prognosis task. Such a phenomenon inspired us to explore how to leverage the already trained supervised and self-supervised models for pathological survival analysis.

Unlike training models on single data modality, multiple modalities are common in medical domains. Medical image data includes radiology images, pathology images, and clinical information like genomic data, etc. Multi-modal learning plays an important role in diagnosis and prognosis as shown in (Chen et al., 2021b). Utilizing both radiology, pathology, and genomic data could provide potential improvement when all modalities are available for all patients. However, there exist missing modality problems when some modality data are missing for certain patients.

1.2.3 Challenge on Energy Efficient Representation Learning for Meta-optics

The rapid developments in deep learning led to the analysis revolution in a number of fields, from autonomous driving to medical image analysis. The advances, however, resulted in requirements of large computational resources, high energy consumption, and longer decision-making time for the deep learning model. The infinite computational requirements of deep learning models would lead to the growth of energy consumption. For the circumstances when large computing resources are not available, the model advances would be limited by the computation requirements.

1.3 Label Efficient Representation Learning on Biological Image

1.3.1 Unsupervised Semantic Segmentation in Microscopy Imaging

Within the CycleGAN framework (Zhu et al., 2017b), many previous studies have tackled unsupervised semantic segmentation in microscopy imaging. Ihle et al. (Ihle et al., 2019b) proposed to use of the CycleGAN framework to segment bright-field images of cell cultures, a live-dead assay of C.Elegans, and X-ray-computed tomography of metallic nanowire meshes. A similar approach was proposed by[28] for facilitating stain-independent supervised and unsupervised segmentation on kidney histology. DeepSynth (Dunn et al., 2019) was proposed to further extend the CycleGAN framework from 2D to 3D nuclear segmentation. Even though the CycleGAN-based unsupervised segmentation approaches have shown decent performance on microscope images, very few studies have investigated the challenging sub-cellular microvilli segmentation with fluorescence microscopy imaging. The sub-cellular microvilli segmentation is challenging due to the highly overlapping and dynamic nature of such small sub-cellular objects (Meenderink et al., 2019; Julio et al., 2008). Different from Pix2Pix GAN (Isola et al., 2017), which requires pixel-level matching between images across two domains, CycleGAN can perform image synthesis without paired images. However, the previous studies emphasized that the macro-level (global distribution level) matching on the number of objects between intensity images and simulated masks improved the segmentation performance (Ihle et al., 2019b). That fact inspired us the question that if the segmentation performance could be further improved by doing more careful matching than the macro-level. To answer the question, we propose a new micro-level matching (mini-batch level) strategy to match the rough number of objects across two domains when training the CycleGAN framework.

1.3.2 Image Synthesis without Annotations

The simple approach to synthesizing new images is to perform image transformations, which include flipping, rotation, resizing, and cropping. Such synthetic images improved the accuracy of image classification upon benchmark datasets (Wang et al., 2022b) by enlarging datasets with synthetic images. Another study (Hore et al., 2015) improved the accuracy of image segmentation (Dice similarity coefficient) with synthetic images by applying data augmentation approaches like random sheering and rotation. A method that is more complex than image transformations is generative adversarial networks (GAN) (Zhu et al., 2017b), which open a new window for synthesizing highly realistic images and have been widely used in different computer vision and biomedical imaging applications. For instance, GAN has synthesized retinal images to map retinal images to binary retinal vessel trees (Costa et al., 2018). The synthetic images can be generated from random noise (Zhang et al., 2018) with geometry constraints (Zhuang and Wang, 2022), and even in high dimensional space (Nimura et al., 2015). To tackle the limitations of needing paired training data requirements, CycleGAN (Isola et al., 2017) was proposed to further advance the GAN technique to broader applications. CycleGAN has shown promise in cross-modality synthesis (Huo et al., 2018a) and microscope image synthesis (Ihle et al., 2019b). DeepSynth (Dunn et al., 2019) demonstrated that CycleGAN can be applied to 3D medical image synthesis.

1.3.3 Microscope Image Segmentation and Tracking

Historically, early approaches utilized intensity-based thresholding to segment a region of interest (ROI) from the background. Ridler et al. (Ridler and Calvard, 1978) use a dynamically updated threshold to segment an object based on the mean intensity of the foreground and the background. Otsu et al. (Otsu, 1979) set a threshold by minimizing the variance of the intraclass. To avoid the sensitivity to all image pixels, Pratt et al. (Pratt, 2007) proposed growing a segmented area from a point, determined by texture similarity. Based on rough annotations, energy functions can be abstracted to segment images by minimizing the aforementioned energy function (Kass et al., 1988). Among such methods, the watershed segmentation approaches are arguably the most widely used methods for intensity-based cell image segmentation (Kornilov and Safonov, 2018).

Object tracking on microscope videos is challenging due to the complex dynamics and vague instance boundaries when at cellular or subcellular resolutions. Gerlich et al. (Gerlich et al., 2003) used optical flow from microscope videos to track cell motion. Ray et al. (Ray and Acton, 2004) tracked leukocytes by computing gradient vectors of cell motions based on active contours. Sato et al. (Sato et al., 1997) designed orientation-selective filters to generate spatial-temporal information by enhancing the motion of cells (de Hauwer et al., 1999). (de Hauwer et al., 1998) also tracked cell motion by applying spatiotemporal analysis on microscope videos. Recent studies have employed machine learning, especially deep learning approaches, for instance, cell segmentation and tracking. Jain et al. (Jain et al., 2007) showed superior performance of a well-trained convolutional network. Baghli et al. (Baghli et al., 2020) achieved 97% prediction accuracy by employing supervised machine learning approaches. To avoid relying on image annotation, Fu et al. (Fu et al., 2022) trained a Convolutional Neural Network without annotation to track large scale fibers in images of material acquired via microscope techniques. However, to the best of our knowledge, no existing studies have investigated the challenging problem of quantifying cellular and subcellular dynamics with pixel-wise instance segmentation and tracking with embedding based deep learning.

1.4 Efficient Feature Representation Learning Methods in Medical Optical Imaging

1.4.1 Unsupervised Representation Learning Model by Contrastive Learning

To extract clinically relevant information from GigaPixel histopathology images is essential in computer-assisted digital pathology (Zhu et al., 2017a; Xu et al., 2015b; Liskowski and Krawiec, 2016). However, pixel-wise annotations are resource extensive given the high resolution of the pathological images. Thus, the fully supervised learning schemes might not be scalable for large-scale studies. Recently, a new family of unsupervised representation learning, called contrastive learning, has shown its superior performance in various vision tasks (Zhuang et al., 2019; Wu et al., 2018; Noroozi and Favaro, 2016; Hjelm et al., 2019). Learning from large-scale unlabeled data, contrastive learning can learn discriminative features for downstream tasks. SimCLR [62] maximizes the similarity between images in the same category and repels the representation of different-category images. Wu et al. (Wu et al., 2018) uses an offline dictionary to store all data representation and randomly select training data to maximize negative pairs. MoCo (He et al., 2020a) introduces a momentum design to maintain a negative sample pool instead of an offline dictionary. Such works demand large batch size to include sufficient negative samples. To eliminate the need for negative samples, BYOL (Grill et al., 2020) was proposed to train a model with an asynchronous momentum encoder. Recently, SimSiam (Chen and He, 2021) proposed to further eliminate the momentum encoder in BYOL, allowing less GPU memory consumption.

1.4.2 Efficient Model Training Adaption on Limited GPU Resources

Limited computing resources are a barrier; thus, multiple research works have been proposed to adapt model training on limited GPU resources. As developed in (Le et al., 2011), (Pal et al., 2019b) model training and data processing can be deployed on multiple GPU devices which enables large batch size and speed up the training process. GPU parallel computing requires supported GPU devices and parallel computing mechanisms. To mitigate the available GPU device limitation, an increased effort has been made to develop memory-efficient training strategies on GPU. Song Han et al.(Han et al., 2016) reduce model size by pruning and Huffman coding. (Luo et al., 2021) proposed to compress the gradient in the training process to reduce the communication load in parallel training. NVIDIA also proposed mixed precision training (Narang et al., 2018) to a half model weight and gradient precision in model training. Besides model size and gradient, ActNN (Chen et al., 2021a) is designed for activation value compression in training.

1.4.3 Efficient Fine-tuning Methods for Pretrained Model on Downstream Task

Supervised pre-trained models (e.g., on ImageNet (Krizhevsky et al., 2017) and BiT (Lu et al., 2022)) have been regarded as a powerful feature extractors and weight initializers in pathological image analysis (Kieffer et al., 2018), (Chen et al., 2022a). However, it is resource-intensive to collect the large-scale annotated images, especially for gigapixel Whole Slide Images (WSIs) (Huo et al., 2021), (David et al., 2019). Without requiring annotations, self-supervised learning (SSL) approaches are leading to a paradigm shift in large-scale pretraining for histopathological image analysis from visual inspection to more accurate quantitative assessment (Yang et al., 2021; Wang et al., 2021c; Ciga et al., 2022; Liu et al., 2021a), with the rapid growth of publicly available large-scale datasets (e.g., The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015b), and Pathology AI Platform (PAIP) (Kim et al., 2021)). In a recent study, Wang et al. (Wang et al., 2021c) utilized the entire TCGA and PAIP dataset to perform a self-supervised pretraining via a vision transformer, called TransPath. TransPath learned the pathological domain-specific information and achieved superior tissue classification performance. Most existing studies focused on developing more powerful pretrained models (Tellez et al., 2018; Mormont et al., 2021), whose resource consumption is increasingly unscalable for academic institutes. Very few, if any, studies have investigated how to take advantage of existing, yet heterogeneous pre-trained models for better performance on downstream tasks. As an example, the pathological data-optimized contrastive learning model TransPath (Wang et al., 2021c) achieved superior performance compared with supervised models (e.g., ImageNet pre-trained) on a classification cohort. Surprisingly, it yielded an inferior performance when it was translated to a cancer prognosis task. Such a phenomenon inspired us to explore how to leverage the already trained supervised and self-supervised models for pathological survival analysis (Zhu et al., 2017d; Tang et al., 2019; Li et al., 2018).

1.4.4 Efficient Multi-Modality Representation Learning on Medical Image

Deep learning-based methods have been successfully applied for automated MSI prediction directly from hematoxylin and eosin (H&E)-stained whole-slide images (WSIs) (Yamashita et al., 2021), (Kather et al., 2019a). Kather et al. (Kather et al., 2019a) developed a ResNet-based model to predict patients with MSI and MSS tumors. Another work (Yamashita et al., 2021) further proposed MSINet and proved the deep learning model exceeded the performance of experienced gastrointestinal pathologists at predicting MSI on WSIs. Despite the vital role of such diagnostic biomarkers (Sidaway, 2020), patients with similar histology profiles can exhibit diverse outcomes and treatment responses. Novel and more specific biomarkers are needed from a whole spectrum of modalities, ranging from radiology (Echle et al., 2021; Wu et al., 2019; Pei et al., 2022a), histology (Wang et al., 2022a; Kather and Calderaro, 2020; Ushizima et al., 2022), and genomics (Braman et al., 2021), (Boehm et al., 2022).

1.5 Energy Efficiency Representation Learning for Meta-optic

The Optical neural network has a high bandwidth (Zhou and Anderson, 1994) and uses light instead of electrical signals to perform matrix multiplications (Duport et al., 2012a), (Larger et al., 2012) which can be much faster and more energy-efficient than traditional digital neural networks. Most optical neural networks (ONN) use a hybrid model structure: implement linear computation with the optic device and non-linear operation digitally (Hughes et al., 2018; De Marinis et al., 2019; Jutamulia and Yu, 1996). Besides the use of optical devices, ONN has been implemented on nanophotonic circuits (Shen et al., 2017), (Fang and Sun, 2015) and light-wave linear diffraction (Lin et al., 2018), (Ovchinnikov et al., 1999) to improve model efficiency. For the non-linear computation, (Miscuglio et al., 2018) have proposed implementing the non-linear operation with the optic device on ONN.

CHAPTER 2

Label Efficient Representation Learning for Biological Optical Imaging

2.1 GAN based Unsupervised Segmentation on Microscopic Image

2.1.1 Introduction

Semantic segmentation is one of the central tasks in microscope image analysis, which segments targeting objects from background context (Wu et al., 1995). Traditionally, the semantic segmentation was performed by unsupervised intensity-based methods, such as watershed (Pinidiyaarachchi and Wählby, 2005), Gaussian mixture model (GMM) (Ragothaman et al., 2016), graph-cut (Leskó et al., 2010) etc. In the past few years, deep learning based methods have been increasingly popular in microscopy imaging, due to their superior accuracy and better generalizability (Moen et al., 2019). However, one of the major limitations in deep learning based semantic segmentation is the need of large-scale annotated images, which is not only tedious, but also resource intensive (Zhang et al., 2017). CycleGAN (Zhu et al., 2017b), a breakthrough generative adversarial network (GAN) (Goodfellow et al., 2014) was proposed recently, which shed light on semantic segmentation with minimal or even no manual annotation (Huo et al., 2018a,b; Zhang et al., 2018c; Chen et al., 2019).

Within the CycleGAN framework (Zhu et al., 2017b), many previous studies have tackled unsupervised semantic segmentation in microscopy imaging. Ihle et al. (Ihle et al., 2019a) proposed to use the CycleGAN framework to segment bright-field images of cell cultures, a live-dead assay of C.Elegans, and X-ray-computed tomography of metallic nanowire meshes. A similar approach was proposed by (Gadermayr et al., 2019) for facilitating stain-independent supervised and unsupervised segmentation on kidney histology. DeepSynth (Dunn et al., 2019) was proposed to further extend the CycleGAN framework from 2D to 3D nuclear segmentation. Even though the CycleGAN-based unsupervised segmentation approaches have shown decent performance on microscope images, very few studies have investigated the challenging sub-cellular microvilli segmentation with fluorescence microscopy imaging. The sub-cellular microvilli segmentation is challenging due to the highly overlapping and dynamic nature of such small sub-cellular objects (Julio et al., 2008; Meenderink et al., 2019).

Different from Pix2Pix GAN (Isola et al., 2017), which requires pixel-level matching between images across two domains, CycleGAN is able to perform image synthesis without paired images. However, the previous studies emphasized that the macro-level (global distribution level) matching on the number of objects between intensity images and simulated masks improved the segmentation performance (Ihle et al., 2019a). That fact inspired us with the question that if the segmentation performance could be further improved by doing more careful matching than the macro-level. To answer the question, we propose a new micro-level matching (mini-batch level) strategy to match the rough number of objects across two domains when training the CycleGAN framework.

In this paper, we develop a deep learning based unsupervised semantic segmentation method for sub-cellular microvilli segmentation using fluorescence microscopy. Meanwhile, we evaluate the performance of micro-level matching strategy, which is enabled by the multi-channel nature of fluorescence images. The contributions of this study are three-fold: (1) We propose the first deep learning based unsupervised sub-cellular microvilli segmentation method; (2) We propose the micro-level matching to ensure the roughly same number of objects across two modalities within each mini-batch, without introducing extra human annotation efforts; (3) Comprehensive analyses are provided to evaluate the outcomes of different augmentation strategies when generating the simulated masks for unsupervised microvilli segmentation.

2.1.2 Methods

Our proposed unsupervised segmentation method consists of two parts: (1) image synthesis, and (2) segmentation. In image synthesis, our goal is to synthesize realistic looking images from the simulated masks. Then, the paired synthetic images and masks are used to train another segmentation network. **Note that, no manual annotations are used in our training either for CycleGAN or U-Net, as an unsupervised framework.**

2.1.2.1 Cycle-consistent image synthesis

The CycleGAN (Zhu et al., 2017b) is used to generate our synthetic training data. As the standard CycleGAN implementation, generators and discriminators are used to transfer the styles between two image modalities. The role of the generators is to convert the real images to another domain, which are typically called ”fake” images. The discriminators then judge if a given image is real or

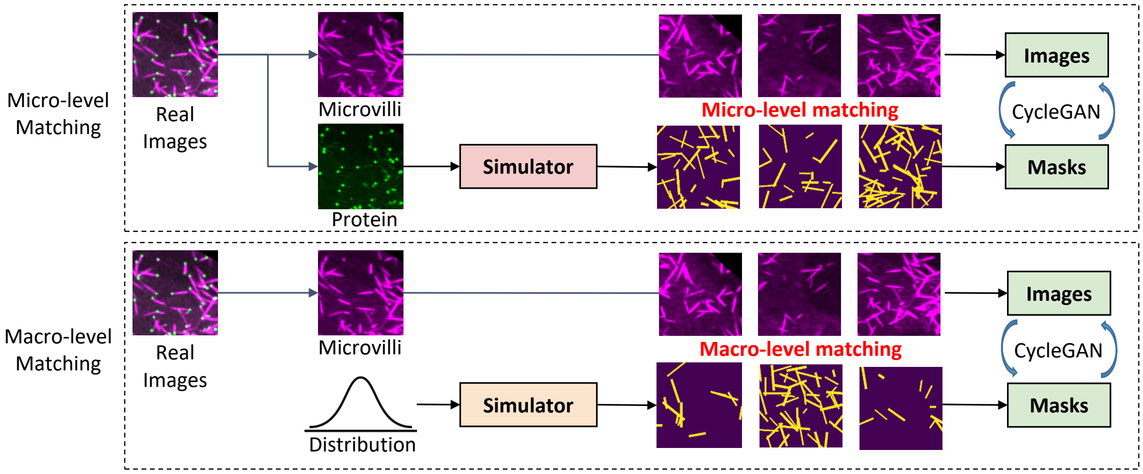


Figure 2.1: This figure shows the general framework of the image synthesis. CycleGAN is used to perform synthesis between real images and simulated masks. In micro-level matching, the green protein channel in fluorescence images is used to achieve the cell counting automatically, which provides the rough numbers of microvilli in the real images. Then, the corresponding simulated masks with the same number of sticks are generated from the simulator when forming a mini-batch for training. In macro-level matching, the numbers of sticks in the simulated masks are randomly generated from the prior global distribution, without mini-batch level correspondence.

fake. The CycleGAN model design creatively forms the entire learning process as a cycle-consistent loop, where the reconstructed fake images after two generators should be close to the original real images. In each branch, the generator tries to generate realistic images, while the discriminator tries to distinguish the fake images from the real ones.

In our unsupervised image segmentation framework (Fig. 2.1), the CycleGAN is employed to synthesize segmentation masks (or called ”annotations”) from the real images I, and synthesize realistic-looking images from simulated segmentation masks M. In the ideal case, the trained generator $G_{I->M}$ can be directly used as a segmentation network to segment new images. However, the quality of synthesis between real images and clean binary masks is typically not optimal since the underlying Poisson distribution of the binary masks is not a realistic distribution in real images (Ihle et al., 2019a). Moreover, the optimization of the KL divergence for training discriminators is more difficult to converge (Gadermayr et al., 2019) using clean binary masks. Therefore, the Gaussian smoothing, random noise, and brightness variations are used to generate augmented masks M_A in addition to the simulated mask images M for better synthetic performance (Fig. 2.2). Then, the trained generator $G_{M_A->I}$ will provide us unlimited fake but realistic-looking images I_F from the

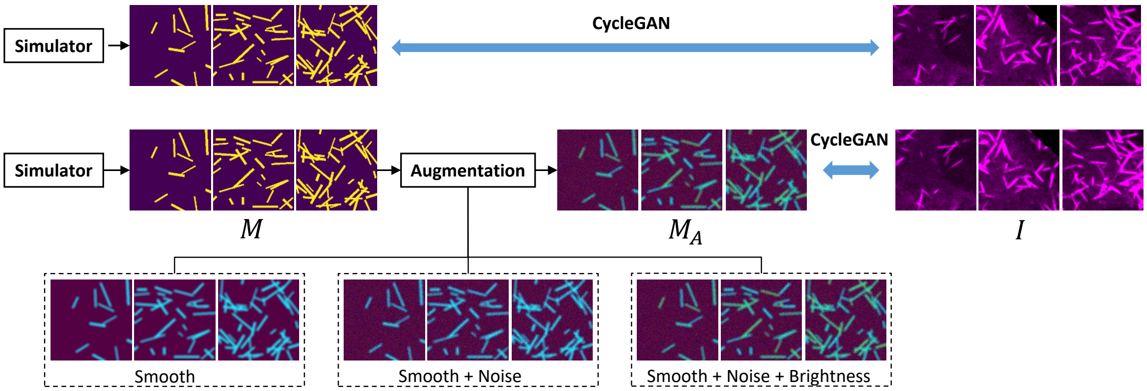


Figure 2.2: This figure demonstrates four experimental designs of providing different augmentations of masks for training the CycleGAN. The first design employs the binary mask directly, while the remaining three designs utilize different augmentation strategies.

simulated and augmented masks M_A. Eventually, the fake images I_F and the clean binary masks M_A (before augmentation) are used to train another independent segmentation network (see **Image Segmentation** section).

2.1.2.2 Micro- and macro-level matching

Compared with traditional pixel-to-pixel conditional GAN design which needs pixel-level correspondence between two modalities, the CycleGAN does not need paired images for training. However, the superior synthetic segmentation performance is typically achieved if the distributions of the number of objects in real image modality and annotation modality are roughly matched (Ihle et al., 2019a; Gadermayr et al., 2019), named as ”macro-level” matching. However, no studies have explored the level of matching in the middle of pixel-level and macro-level. In this study, we proposed the idea called ”micro-level” matching, which matches the number of objects in each mini-batch (Fig. 2.1). For example, if a real image has roughly 21 microvilli, we will provide a simulated mask with the same 21 sticks, when forming the mini-batch. Then, the next question is how can we get the rough number of objects from the real images. In this study, we utilize the multichannel nature of fluorescence microscopy to split the microvilli marker mCherry-Espin (magenta color objects) and microvilli tip marker EGFP-EPS8 (green color objects). Using the simple intensity thresholding-based cell counting algorithm (Refai et al., 2003), the rough number of protein objects is easily achieved. The numbers are then used as the rough number of microvilli to simulate

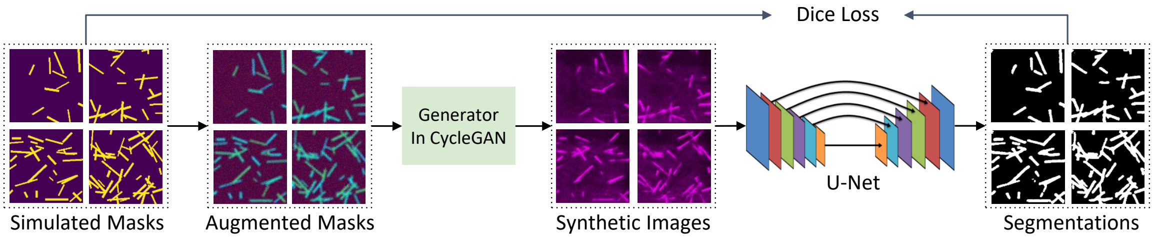


Figure 2.3: This figure shows the segmentation pipeline. The input images of the U-Net model are the synthetic fake images from trained CycleGAN's generator, while the annotations are the simulated masks. Note that, no manual annotations are used in our training either for CycleGAN or U-Net, as an unsupervised framework.

the corresponding mask files with the same number of objects, as the micro-level matching. Note that we only match the number of objects in the micro-level matching, where the spatial distribution of the objects is still random.

2.1.2.3 Image segmentation

U-Net (Ronneberger et al., 2015) is employed as the segmentation backbone network, which is a fully convolutional neural network and is widely used in image segmentation tasks. The segmentation part of our framework is shown in Fig. 2.3. In our proposed unsupervised segmentation framework, the input images of U-Net are the fake microvilli images, which are generated from the simulated masks using $G_{M->I}$ or $G_{M_A->I}$ from trained CycleGAN. Then the Dice loss function is calculated by comparing the predicted segmentation with the simulated binary masks. The traditional deep neural network typically needs a large number of annotated images to train a segmentation network. Using our design, however, we can generate an unlimited number of training data to train the segmentation network without any manual annotation efforts.

2.1.3 Data and experimental design

2.1.3.1 Microvilli images

Twelve microvilli images acquired using fluorescence microscopy were used as training data, where each image had $\approx 900 \times 900$ pixels with pixel resolution 1.1 μm. Then, 500 image patches with 128×128 pixels were randomly sampled from the twelve images to train the CycleGAN as the real images. Then, another independent microvilli video with 20 frames was used as testing data to evaluate the performance of the proposed unsupervised segmentation methods. Each frame has 256×256 pixels with pixel resolution 1.1 μm. All microvilli in each frame were densely annotated manually by an experienced biologist as the gold standard segmentation.

2.1.3.2 Experimental design

In order to test if micro-level matching can improve the unsupervised segmentation performance, we performed experiments using both macro-level and micro-level matching. For micro-level matching, the number of sticks of each image was obtained by automatically counting the number of green proteins. For macro-level matching, the number of sticks for each image was randomly sampled from a uniform distribution (range from 11 to 63), according to the distribution of proteins.

As shown in Fig 2.2, we have four different augmentation settings to generate the simulated masks in annotation domain:

Binary masks: The binary masks were directly simulated as the images in the annotation domain, without any augmentation. Based on (Meenderink et al., 2019) and the prior biological knowledge of microvilli, the width of each microvilli was simulated between 2 to 5 μm, while the length was simulated between 10 to 50 μm. As the pixel resolution of all our images was 1.1 μm, we randomly generated sticks with 2 to 4 pixels width and 9 to 45 pixels length from a uniform distribution.

Gaussian smoothing: The first augmentation was Gaussian smoothing, where a Gaussian filter with a kernel size of 5×5 was applied to the binary masks.

Random noise: Upon the Gaussian smoothing, the random Gaussian noise was further applied to the entire mask image. The values of random noise ranged from 0 to 255 following Gaussian distribution.

Different brightness: To further introduce the global intensity variations, random intensity values (200 to 255) were assigned to each stick in binary masks, where the maximum foreground intensity value was 255.

To improve the segmentation performance, CycleGAN was employed to synthesize cell images for U-Net model training. In our experiment, CycleGAN is used to learn the mapping from simulated masks to real microvilli cell images. We built up a dataset in these two domains as CycleGAN model's input. Our CycleGAN model was trained for 60 epochs. According to the training loss, the generator trained for 50 epochs shows the best performance. The generator trained in CycleGAN

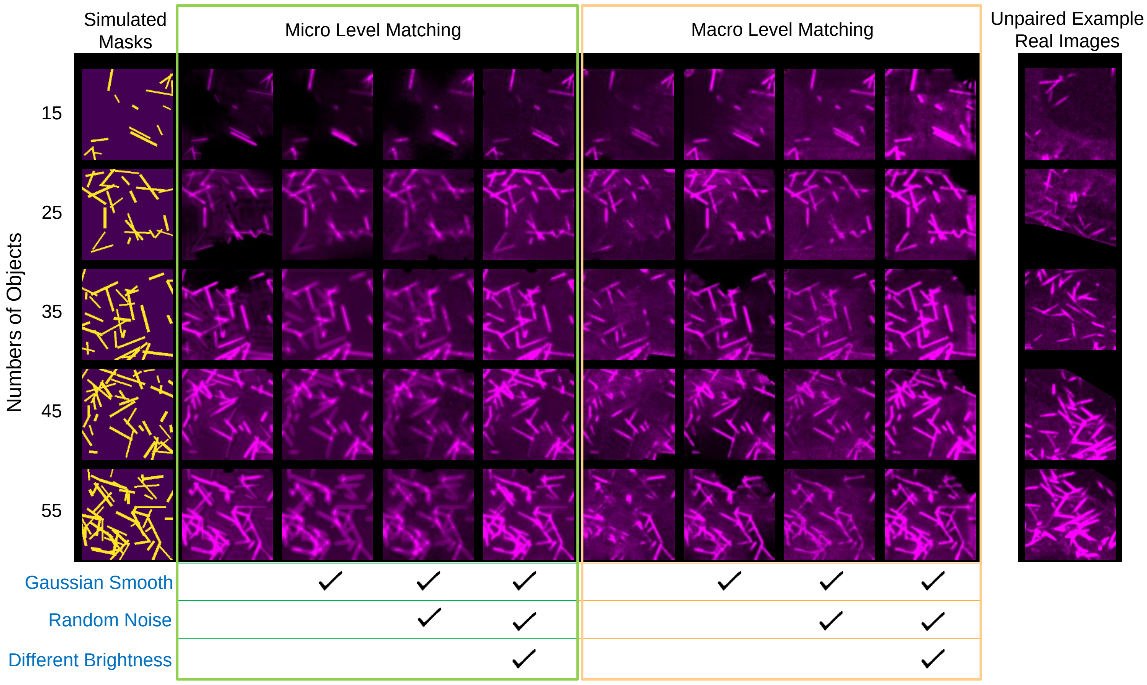


Figure 2.4: The synthesis results of different experimental designs are provided in this figure. The first column is the initial simulated masks with different numbers of objects (sticks). The middle columns exhibit the synthetic images from the masks using different augmentation strategies. The last column is five randomly selected real images, which are unpaired to the masks in CycleGAN framework.

will be used to synthesize microvilli cell images based on simulated mask images.

CycleGAN model cannot cover all details using original frames as input which has too many cells. For both CycleGAN and U-Net, the input images were all with 128×128 resolution cropped from original frames and then resized to 256×256 during training. When applying trained U-Net on testing microvilli images, each testing image was first split into four 128×128 images, and the final segmentation was achieved by concatenating the corresponding four predictions back to the original resolution. The Dice results were calculated in the original 256×256 resolution for testing images. The CycleGAN and U-Net were deployed on a computer with a GeForce GTX 1060 Graphic Card with 6 GB memory. To get better synthesised data and avoid over-fitting, the CycleGAN was trained with 50 epochs and the U-Net was trained with 10 epochs for all experiments. According to the prediction performance, U-Net has the best performance after 10 epochs. The results from the last epochs were reported in this paper.

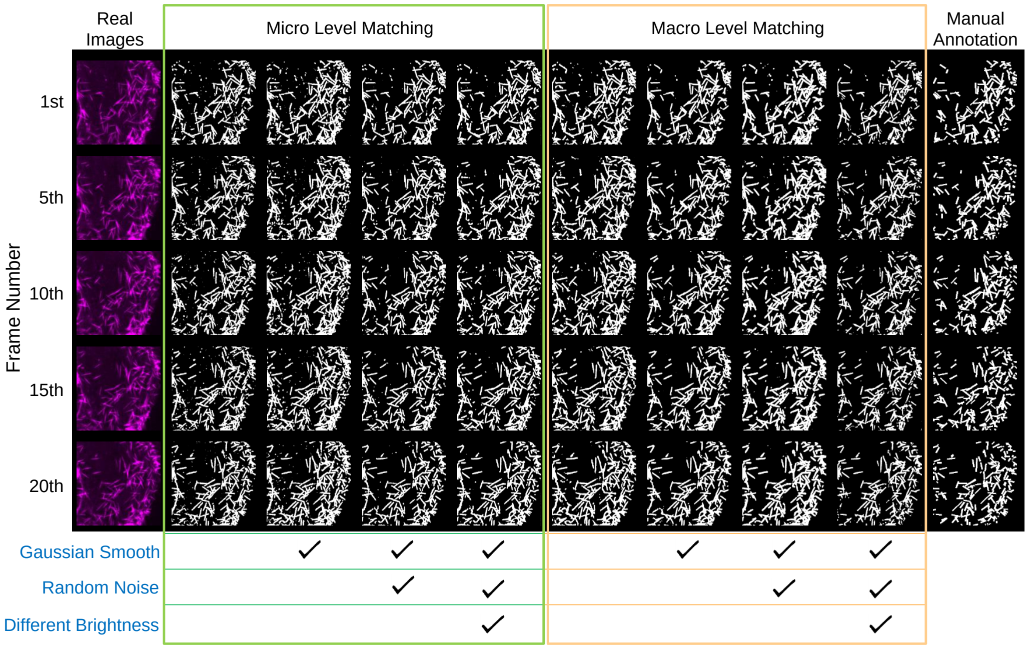


Figure 2.5: The final segmentation results from U-Net are presented in this figure. Each row shows the segmentation results at different frames in a microvilli video.

Table 2.1: The average Dice values of different experiments.

Exp.	Smooth	Noise	Bright.	$D_{(w=1)}$	$D_{(w=2)}$	$D_{(w=3)}$	$D_{(w=4)}$	$D_{(w=5)}$
Micro-level matching				0.3818	0.4860	0.5459	0.5605	0.5628
	✓			0.3650	0.4691	0.5301	0.5535	0.5667
	✓	✓		0.3738	0.4783	0.5367	0.5511	0.5547
	✓	✓	✓	0.3810	0.4865	0.5479	0.5650	0.5730
Macro-level matching				0.3639	0.4717	0.5364	0.5583	0.5675
	✓			0.3811	0.4918	0.5557	0.5776	0.5888
	✓	✓		**0.3902**	**0.4981**	**0.5607**	**0.5965**	**0.6169**
	✓	✓	✓	0.3894	0.4903	0.5467	0.5615	0.5631

"*D*" indicate the Dice score, *w* means the width of the ground truth.

2.1.4 Results

Considering both micro- and macro-level matching with different augmentation strategies, we performed eight experiments by training eight different CycleGAN networks. The qualitative results of image synthesis from eight different CycleGAN networks are provided in Fig. 2.4.

Then, the synthetic images were used to train eight different U-Net models using synthetic training image patches and applied to the real testing images. For testing images, the manual annotation was performed by tracking center line fragments of each microvillus (annotated by the experienced biologist) since the traditional contour-based annotations were extremely difficult on the tiny subcellular structures. To evaluate the segmentation results, we assigned different widths to the manual segmentation and reported the results in Table. 2.1. The corresponding qualitative results of segmentation are provided in Fig. 2.5. According to the microvilli cell's biological characteristics, manual annotation images are presented with width=3. From the results, the macro-level matching with Gaussian smoothing and random noise achieved the best performance across different widths of manual annotation. The micro-level matching did not improve the segmentation performance. Micro-level pairing can achieve higher accuracy on training datasets because its pairing is detailed to fit training dataset properties. Macro-level pairing is more robust. U-Net model trained by macro-level pairing performs better than micro-level pairing on the new dataset. The standard Dice similarity coefficient metrics were used to evaluate different methods. The video of microvilli frames and our unsupervised segmentation results are presented in the **supplementary materials:** https://github.com/iamliuquan/GAN_based_segmentation.

2.1.5 Conclusion

In this study, we proposed the first deep learning solution to enable unsupervised sub-cellular microvilli segmentation. Beyond the current standard macro-level matching strategy, we utilized the multi-channel nature of fluorescence microscopy to enable the micro-level matching of the number of objects in each mini-batch without introducing new human annotation efforts. From the experimental results, we conclude that the micro-level matching of object numbers at the mini-batch level did not lead to better segmentation performance. From the comprehensive analyses of introducing noise, smoothness and brightness, the Gaussian smoothing and random noise on the simulated annotations with macro-level matching resulted in the best microvilli segmentation performance.

2.2 Annotation-free Synthetic Instance Segmentation and Tracking for Microscope Video

2.2.1 Introduction

Capturing cellular and subcellular dynamics through microscopy approaches helps domain experts in characterizing biological processes (Meenderink et al., 2019) in a quantitative manner, leading to advanced biomedical applications (e.g., drug discovery) (Arbelle et al., 2018).

Numerous image processing approaches have been proposed for precise instance object segmentation and tracking. Most of the previous solutions (Al-Kofahi et al., 2018; Korfhage et al., 2020; Van Valen et al., 2016) follow a similar “two-stage” strategy: I. segmentation on each frame, and II. frame-by-frame association across the video. In recent years, a new family of “single-stage” algorithms was enabled by cutting-edge pixel-embedding based deep learning (Zhao et al., 2020a; Payer et al., 2018). Such methods enforce the spatiotemporally consistent pixel-wise feature embedding for the same cellular or subcellular objects across video frames. However, pixel-wise annotations require spatial (segmentation) and temporal (tracking) consistency. Such labeling efforts are typically expensive, and potentially unscalable, for microscope videos due to I. dense objects (e.g., overlapping or touching), and II. high dynamics (e.g., irregular motion and mitosis). Therefore, better learning strategies are desired beyond the current human annotation based supervised learning.

Adversarial simulation has provided a scalable option to create realistic synthetic environments without extensive human annotations. Particularly striking examples include a) using computer games such as Grand Theft Auto to train self-driving deep learning models (Johnson-Roberson et al., 2016), b) using a simulation environment Gazebo to train robotics (Zamora et al., 2016), and c) using a SUMO simulator to train traffic management artificial intelligence (AI) (Kheterpal et al., 2018).

In this paper, we propose an annotation-free synthetic instance segmentation and tracking (ASIST) method with adversarial simulation and single-stage pixel-embedding based learning. Briefly, the ASIST framework consists of three major steps: I. unsupervised image-annotation synthesis, II. video and temporal annotation synthesis, and III. pixel-embedding based instance segmentation and tracking. As opposed to traditional manual annotation-based pixel embedding deep learning, the proposed ASIST method is annotation-free (Figure.4.1).

To achieve the annotation-free solution, we simulated cellular or subcellular structures with three important aspects: shape, appearance and dynamics (Fig.2.7). To evaluate our proposed

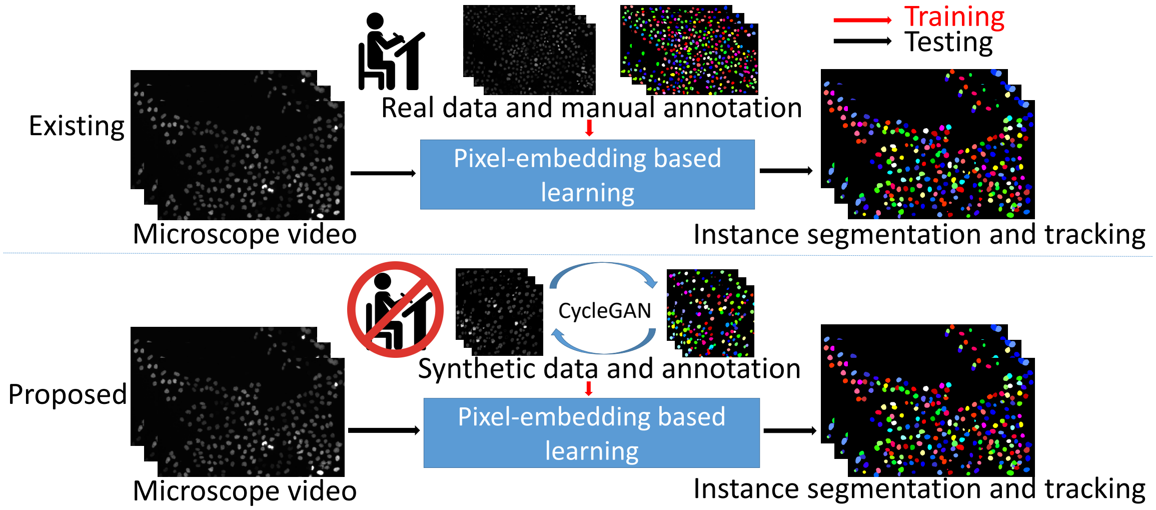

Training
Testing
Existing
Real data and manual annotation
Pixel-embedding based learning
Microscope video
Instance segmentation and tracking
Proposed
CycleGAN
Synthetic data and annotation
Pixel-embedding based learning
Microscope video
Instance segmentation and tracking

Figure 2.6: The upper panel shows the existing pixel-embedding deep learning based single-stage instance segmentation and tracking method, which is trained by real microscope video and manual annotations. The lower panel presents our pro-posed annotation-free ASIST method, with synthesized data and annotations from adversarial simulations.

ASIST method, microscope videos of both cellular (i.e., HeLa cell videos from ISBI Cell Tracking Challenge (Maška et al., 2014; Ulman et al., 2017)) and subcellular (i.e., microvilli videos from in house data) objects were included in this study. The HeLa cell videos have larger shape variations compared with microvilli videos. From the results, our ASIST method achieved promising accuracy compared with fully supervised approaches.

In summary, this paper has three major contributions:

- We propose the ASIST annotation-free framework, aggregating adversarial simulations and single-stage pixel embedding based deep learning.

- We propose a novel annotation refinement approach to simulate shape variations of cellular objects, with circles as a middle representation.

- To our best knowledge, our proposed approach is the first annotation-free solution for single-stage pixel embedding deep learning based cell instance segmentation and tracking.

This research was supported by Vanderbilt Cellular, Biochemical and Molecular Sciences Training Grant 5T32GM008554-25, the NIH NIDDK National Research Service Award F31DK122692, NIH Grant R01-DK111949 and R01-DK095811.

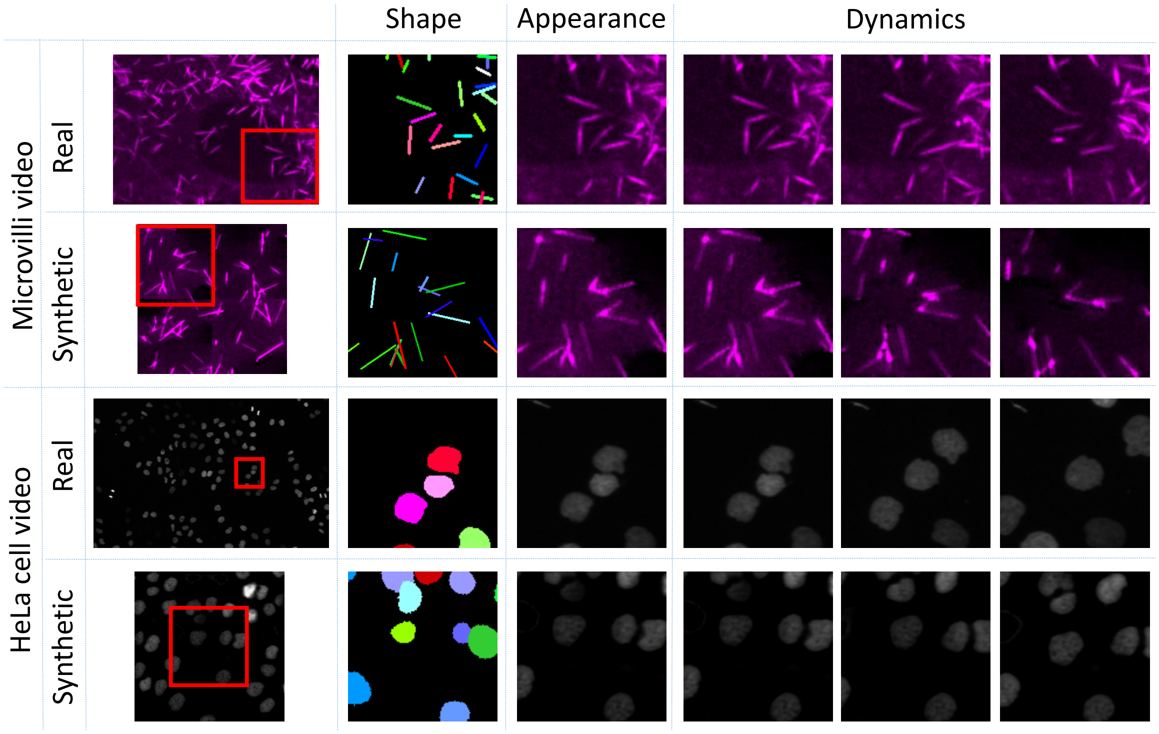


Figure 2.7: Real and synthetic video of Hela cell and microvilli consisting of three aspects: shape, appearance and dynamics. The ”shape” is defined as the underlying shape of the manual annotations. The ”appearance” is defined by the various appearances of objects. The ”dynamics” indicates the mitigation of cellular and subcellular objects.

2.2.2 Related Work

2.2.2.1 Image synthesis

The simplest approach to synthesize new images is to perform image transformations, which includes flipping, rotation, resizing, and cropping. Such synthetic images improved the accuracy of image quantification upon benchmark datasets (Simard et al., 2003) by enlarging them with synthetic images. Another study (Drozdzal et al., 2018) synthesized new images by applying data augmentation approaches like random sheering and rotations to training data.

A method that is more complex than image transformations are generative adversarial networks (GAN) (Goodfellow et al., 2014), which open a new window of synthesizing highly realistic images, and have been widely used in different computer vision and biomedical imaging applications. For instance, GAN has synthesized retinal images to map retinal images to binary retinal vessel trees (Costa et al., 2017). The synthetic images can be generated from random noise (Zhang et al., 2018) with geometry constraints (Zhuang and Wang, 2020), and even in high dimensional space (Liu et al., 2018). To tackle the limitations of needing paired training data requirements, CycleGAN (Zhu et al., 2017b) was proposed to further advance the GAN technique to broader applications. CycleGAN has shown promise in cross-modality synthesis (Huo et al., 2018a) and microscope image synthesis (Ihle et al., 2019b). DeepSynth (Dunn et al., 2019) demonstrated that CycleGAN can be applied to 3D medical image synthesis.

2.2.2.2 Microscope image segmentation and tracking

Historically, early approaches utilized intensity-based thresholding to segment a region of interest (ROI) from the background. Ridler et al. (Ridler and Calvard, 1978) use a dynamic updated threshold to segment an object based on the mean intensity of the foreground and the background. Otsu et al. (Otsu, 1979) set a threshold by minimizing variance of the intraclass. To avoid the sensitivity to all image pixels, Pratt et al. (Pratt, 2007) proposed growing a segmented area from a point, determined by texture similarity. Based on rough annotations, energy functions can be abstracted to segment images by minimizing the aforementioned energy function (Kass et al., 1988). Among such methods, the watershed segmentation approaches are arguably the most widely used methods for intensity based cell image segmentation (Kornilov and Safonov, 2018)..

Object tracking on microscope videos is challenging due to the complex dynamics and vague instance boundaries when at cellular or subcellular resolutions. Gerlich et al. (Gerlich et al., 2003) used optical flow from microscope videos to track cell motion. Ray et al. (Ray and Acton, 2004) tracked leukocytes by computing gradient vectors of cell motions based on active contours. Sato et al. (Sato et al., 1997) designed orientation-selective filters to generate spatio-temporal information by enhancing the motion of cells. (de Hauwer et al., 1998, 1999) also tracked cell motion by applying spatiotemporal analysis on microscope videos.

Recent studies have employed machine learning, especially deep learning approaches, for instance cell segmentation and tracking. Jain et al. (Jain et al., 2007) showed superior performance of a well-trained convolutional network. Baghli et al. (Baghli et al., 2020) achieved 97% prediction accuracy by employing supervised machine learning approaches. To avoid relying on image annotation, Yu et al. (Yu et al., 2018) trained a Convolutional Neural Network without annotation to track large scale fibers in images of material acquired via microscope techniques. However, to the best of our knowledge, no existing studies have investigated the challenging problem of quantifying cellular and subcellular dynamics with pixel-wise instance segmentation and tracking with embedding based deep learning.

2.2.3 Methods

Our study has three steps: unsupervised image-annotation synthesis, video synthesis and instance segmentation and tracking (Fig.4.3).

2.2.3.1 Unsupervised image-annotation synthesis

The first step is to train a CycleGAN based approach (Zhu et al., 2017c) to directly synthesize annotations from microscope images, and vice versa. Compared with the tasks in computer vision, the objects in microscope images are often repetitive with more homogeneous shapes. Therefore, with knowledge of shapes associated with microvilli (stick-shaped) and HeLa cell nuclei (ball-shaped), we randomly generate fake annotations with repetitive sticks and circles to model the shape of microvilli and HeLa cells, respectively. When we train the CycleGAN on microvilli images, we clean the green marks on raw microvilli images which is EPS8 protein by splitting channel of RGB images. The network structure, training process and parameters follows (Liu et al., 2020). The generator in CycleGAN consists of an encoder, transformer and decoder. We used ResNet (He

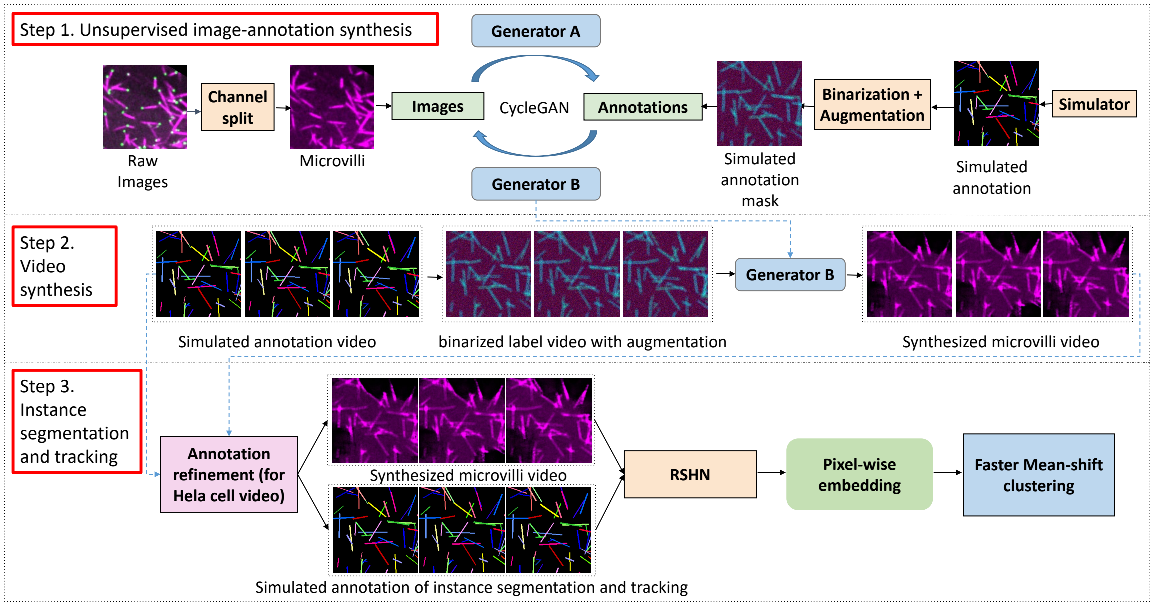

Step 1. Unsupervised image-annotation synthesis
Generator A
Channel split
Raw Images
Microvilli
Images
CycleGAN
Annotations
Generator B
Simulated annotation mask
Binarization + Augmentation
Simulated annotation
Simulator
Step 2. Video synthesis
Simulated annotation video
binarized label video with augmentation
Generator B
Synthesized microvilli video
Step 3. Instance segmentation and tracking
Annotation refinement (for Hela cell video)
Synthesized microvilli video
Simulated annotation of instance segmentation and tracking
RSHN
Pixel-wise embedding
Faster Mean-shift clustering

Figure 2.8: This figure shows the proposed ASIST method. First, CycleGAN based image-annotation synthesis is trained using real microscope images and simulated annotations. Second, synthesized microscope videos are generated from simulated annotation videos. Last, an embedding based instance segmentation and tracking algorithm is trained using synthetic training data. For HeLa cell videos, a new annotation refinement step is introduced to capture the larger shape variations.

et al., 2016) with 9 residual blocks as the encoder in both Generator A and Generator B in the deep learning architecture. We have tried to employ U-Net (Ronneberger et al., 2015) as the encoder as well, suggested by (Liu et al., 2020). Based on the our experience, ResNet generally has superior performance compared with U-Net. As a result, the ResNet is employed as the generator through all experiments in this paper.

2.2.3.2 Video synthesis

Using an annotation-to-image generator (marked as Generator B) from the above CycleGAN model, synthetic intensity images can be generated from simulated annotations. Since a video dataset represents a compilation of image frames, we extend the utilization of the trained Generator B from “annotation-to-image” to ”annotation frames-to-video”. Briefly, simulated annotation videos are generated by our annotation simulator with variations in shape and dynamics. Then, each annotation video frame is used to generate a synthetic microscope image frame. After repeating such a process for the entire simulated annotation videos, synthetic microscope video is achieved for microvilli and

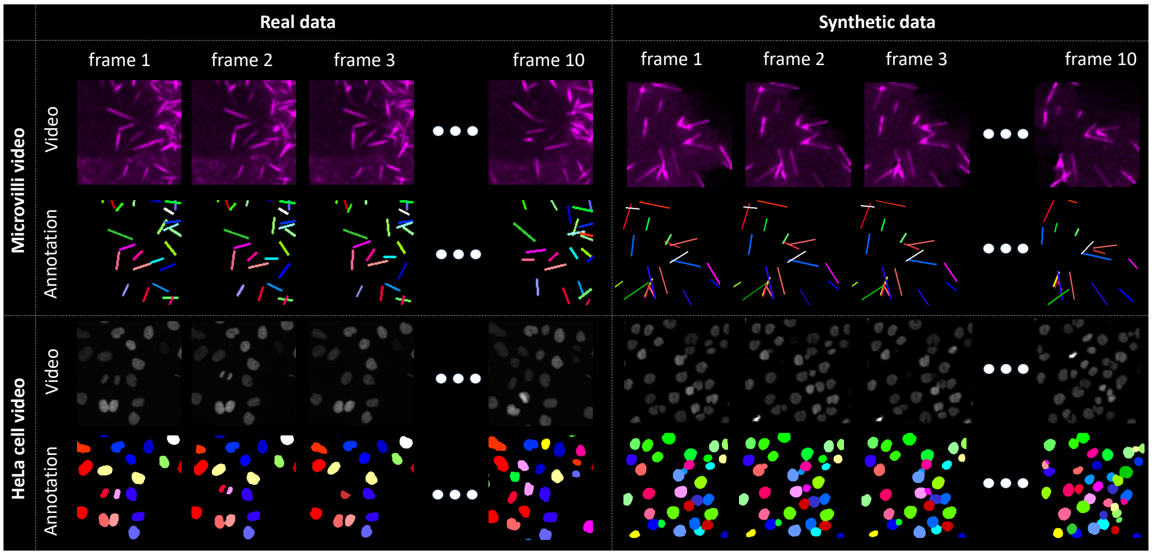


Figure 2.9: The left panel shows real microscope videos as well as manual annotations. The right panel presents our synthetic videos and simulated annotations.

HeLa cells, respectively.

2.2.3.2.1 Microvilli simulation

As shown in Fig.2.9, we model the shape of microvilli as sticks (narrow rectangles) to simulate microvilli videos. The simulated microvilli annotation videos are determined by the following operations:

Object number: Different numbers of objects are evaluated when simulating microvilli videos. The details are presented in §**Experimental design**.

Translation: Instance annotations are translated by 1 pixel at 50% probability.

Rotation: Each instance label is randomly rotated by 1 degree at 50% probability.

Shortening/Lengthening: Each object has 50% probability to become longer or shorter by 1 pixel. Each object can only become longer or shorter across the video.

Moving in/out: To simulate the instance moving in and out from the video scope, we generate frames in larger size (550 × 550 pixels) and center-cropped into the target size (512 × 512 pixels).

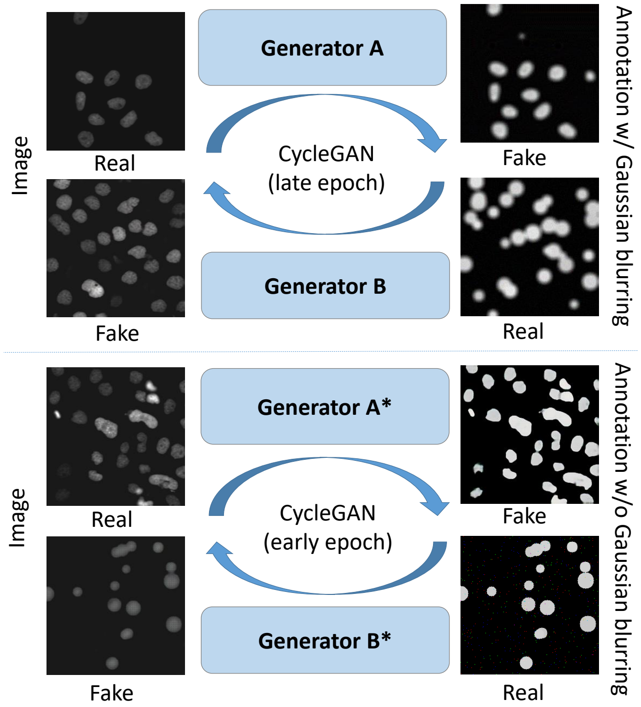


Figure 2.10: The upper panel shows the CycleGAN that is trained by real images and simulated annotations with Gaussian blurring. The lower panel shows the CycleGAN that is trained by the same data without Gaussian blurring. The Generator B is used to generate synthetic videos with larger shape variations from circle representations, while the Generator A* generate sharp segmentation for the annotation registrations.

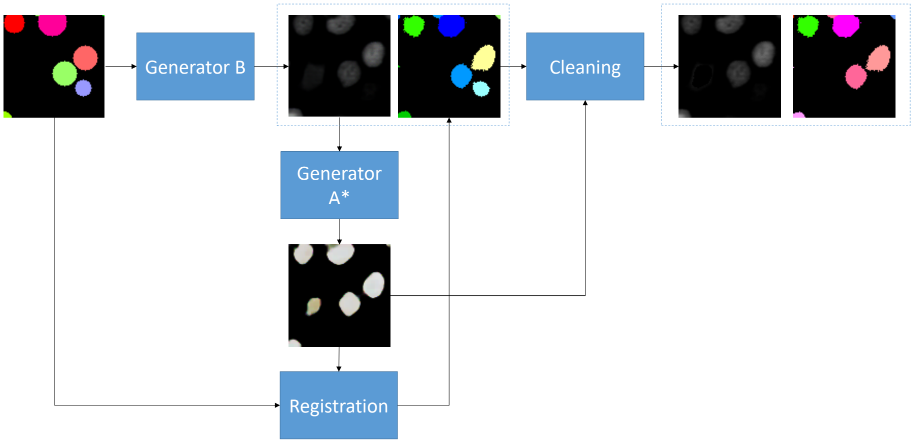


Figure 2.11: This figures shows the workflow of the annotation refinement approach. The simulated circle annotations are fed into Generator B to synthesize cell images. We used Generator A* in Fig.2.10 to generate sharp binary masks from synthetic images. Then, we registered simulated circle annotations to binary masks to match the shape of cells in synthetic images. Last, an annotation cleaning step was introduced to delete the inconsistent annotations between deformed instance object masks and binary masks.

2.2.3.2.2 HeLa cell simulation

The HeLa cells have higher degrees of freedom in terms of shape variations, compared with microvilli. In this study, we proposed an annotation refinement strategy, to generate shape consistent synthetic HeLa cell videos and annotations, using circles as middle representations (Fig. 2.10), without introducing manual annotations. The simulated videos and annotations of HeLa cells are determined by the following operations:

Object number:The numbers of objects are evaluated when simulating HeLa cell videos. The details are presented in §**Experimental design**.

Translation: The instance annotation center can be moved by N pixels. N will be described in §**Experimental design**.

Radius changing: Radius of annotations has 10% probability to get bigger or smaller by 1 pixel.

Disappearing: Existing instance cells are randomly deleted from certain frames in videos.

Appearing: New instance cells shows up from certain frame in videos randomly. New cells will be added to the video from the appearing frame.

Mitosis: Mitosis is the process of cell replication and splitting. To simulate HeLa cell mitosis, we randomly define ”mother cells” at the nth frame. At the $n+1$th frame, we delete the ”mother cells” and randomly create two new cells nearby. Based on biological knowledge, these two new instances are typically smaller than normal instances, and will grow up bigger and move randomly like other instance annotations.

Overlapping: We allow partial overlap between cells. The minimum distance between two cells are set to be 70% of the total diameter between two cells.

Size change: The radius of instance annotation has a 10% probability to become larger by 1 pixel or become smaller by 1 pixel.

2.2.3.3 Annotation refinement for HeLa cell video simulation

After training the initial CycleGAN synthesis, we are able to build simulated videos (with circle representation) as well as their corresponding synthetic microscope videos. However, circles are not the exact shape of annotations for synthetic videos. To further achieve consistent synthetic videos and annotations, we proposed an annotation refinement framework, which has a workflow shown in Fig. 2.11.

2.2.3.3.1 Binary mask generation

We trained CycleGAN to generate a binary mask of synthetic cell images. Unique from CycleGAN in §**Unsupervised image-annotation synthesis**, we used training data without applying Gaussian blurring and used the model from an early epoch. From our experiments, we observed that the early epochs of the CycleGAN training focused more on intensity adaptations rather than shape adaptations. The trained Generator A is used to generate sharp binary masks as templates in the following annotation registration step.

2.2.3.3.2 Annotation deformation (AD)

To bridge the gap between circle representations and HeLa cell shape annotations, a non-rigid registration approach from ANTs (Avants et al., 2011) is used to deform the circle shapes to the HeLa cell shapes. Briefly, we used generator B to synthesize cell images based on our simulated annotations. In the mask generation, we used generator A* to generate binary masks and registered the circle shape annotations to the binary masks. In that case, we keep the label numbers of circle representations, and deform their shapes to fit the synthetic cells.

2.2.3.3.3 Annotation cleaning (AC)

When performing image-annotation synthesis using CycleGAN, it is very likely to have a slightly different number of objects between HeLa cell images and annotations without using paired training data. To make the synthetic videos and simulated annotations to have more consistent numbers of objects, we introduce an annotation cleaning step (Fig. 2.11). First, we generate binary masks of simulated images using the Generator A*. Second, we clean up the inconsistent objects and annotations by comparing deformed simulated annotations and binary masks. Briefly, pseudovideos are simulated annotations. instance annotations are achieved from binary masks, by treating any connected components as instances. Third, if an instance object in the deformed simulated annotations is not 90% covered by binary masks, we re-assign the label as background. On the other hand, if a pseudo instance object from the binary masks is not 90% covered by deformed simulated annotations, we re-assign the corresponding region in the intensity image with the average background intensity values. In sum, the consistent synthetic videos and deformed simulated instance annotations are achieved with annotation cleaning (Fig. 2.11).

2.2.3.4 Instance segmentation and tracking

From the above stages, the synthetic videos and corresponding annotations are achieved frame-by-frame. The next step is to train our instance segmentation and tracking model. We used the recurrent stacked hourglass network (RSHN) (Payer et al., 2018) as the instance segmentation and tracking backbone to encode the embedding vectors of each pixel. The RSHN is a stacked hourglass network with a convolutional gated recurrent unit to process temporal information. The ideal pixel-embedding has two properties: (1) embedding of pixels belonging to the same objects should be similar across the entire video, and (2) the embedding of pixels belonging to different objects should be different. For a testing video, we employed the Faster Mean-shift algorithm (Zhao et al., 2020a) to cluster pixels to objects as the instance segmentation and tracking results. The embedding-based deep learning methods approach the instance segmentation and tracking as a ”single-stage” approach, which is a simple and generalizable solution across different applications (Payer et al., 2018; Zhao et al., 2020a).

2.2.4 Experimental design

2.2.4.1 Instance segmentation and tracking on microvilli video

2.2.4.1.1 Data

Two microvilli videos captured by fluorescence microscopy are in 1.1μm pixel resolution. Training data is one microvilli video in 512×512 in pixel resolution. Testing data is another microvilli video in the size of 328×238 pixels. Due to the heavy load of manual annotations on video frames, we only annotated the first ten frames of both videos as the golden standard. The annotation work includes two parts: 1) first we annotated each microvilli structure including overlapping or densely distributed areas; 2) secondly, each instance has been assigned consistent labels across all frames in same video. The manual annotation labor on both training and testing data takes roughly a week of work from a graduate student. This long manual annotation process shows the value of annotation-free solutions in quantifying cellular and subcellular dynamics.

2.2.4.1.2 Experimental design

In order to assess the performance of our annotation-free instance segmentation and tracking model, the proposed method is compared with the model trained with manual annotations on the same testing microvilli video. The different experimental settings are shown as the following:

Self: The testing video with manual annotations was used as both training and testing data.

Real: Another real microvilli video with manual annotations were used as training data.

Microvilli-1: One simulated video which consisted of 100 instances in size of 512×512 pixels was used as training data. The ”Microvilli-1 10 frames” indicated only 10 frames were used, while other simulated data used 50 frames.

Microvilli-5: Five simulated videos with 512×512 pixel resolutions were used as training data. The number of objects were empirically chosen to be between 80 to 220.

Microvilli-20: We further spatially split each 512×512 video in Microvilli-5 to four 256×256 videos to form a total of 20 simulated videos with half resolution.

2.2.4.2 Instance segmentation and tracking on HeLa cell video

2.2.4.2.1 Data

HeLa cell videos (N2DL-HeLa) were obtained from the ISBI Cell Tracking Challenge (Maška et al., 2014; Ulman et al., 2017). The cohort has two 92-frame HeLa cell videos in size of 1100×700 pixels with annotations. The second video with complete manual annotations is used as the testing data for all experiments.

2.2.4.2.2 Experimental design

For experiments using an annotation-free framework, synthetic videos and simulated annotations are used for training. As a comparison experiment, experiments trained with annotated data used two N2DL-HeLa videos with annotations as training data. Our experiment settings are described as follows:

Self: The testing video with manual annotations was used as both training and testing data. The patch size of 256×256 was used, following (Payer et al., 2018; Zhao et al., 2020a).

Self-HW: The testing video with manual annotations was used as both training and testing data. The patch size of 128×128 was used, as a half window (HW) size.

HeLa: Our training data was 10 simulated videos with 512×512 resolution containing approximately 150 objects, including 20 cell appearing events, 20 cell disappearing events, and 5 or 10 mitosis events. The numbers were empirically chosen. This experiment employed the circle annotations directly as the baseline performance. The patch size of 256×256 was used.

HeLa-AD: The above simulated data was used for training, with an extra annotation deformation (AD) step.

HeLa-AD+AC: The above simulated data was used for training, with extra AD and annotation cleaning (AC) steps.

HeLa-AD+AC+HW: The above simulated data was used for training, with extra AD and AC steps. The patch size of 128×128 was used, as a half window (HW) size.

2.2.4.3 Evaluation matrix

TRA, DET, and SEG are the standard metrics in the ISBI cell tracking challenge (Matula et al., 2015), evaluating the performance of tracking, detection and segmentation, respectively. The ISBI Cell Tracking Challenge used these three metrics as *de facto* standard measurements based on Acyclic Oriented Graph Matching (AOGM) algorithms. The instance objects are presented as the nodes of the acyclic oriented graphs, while the tracking results are modeled as the vertices of the graphs. Then, graphs are obtained from both ground truth annotations and the predicted results to evaluate the accuracy of detection (DET) and tracking (TRA). SEG evaluates the overlap of predicted objects with true objects. The TRA, DET and SEG range from 0 to 1, where 0 and 1 indicate the worst and best performance, respectively. The details of such metrics can be found in (Matula et al., 2015).

2.2.5 Results

2.2.5.1 Instance segmentation and tracking on microvilli videos

The qualitative and quantitative results are presented in Fig. 2.12 and Table. 2.2. From the quantitative results shown in Table. 2.2, the best performance according to the evaluation metric scores was achieved by Microvilli-20 without using manual annotations. By contrast,it took one week of manual annotation labor from a graduate student to annotate only 10 frames of RSHN (Self) and RSHN (Real). One salient feature of achieving better performance of the proposed framework is the larger number of total simulated training video.

Table 2.2: DET, SET and TRA values of different experiments on microvilli video.

Exp.	T.V.	T.F.	DET	SEG	TRA
RSHN (Self)	1	10	0.662	0.298	0.629
RSHN (Real)	1	10	0.357	0.169	0.334
ASIST (Microvilli-1)	1	10	0.580	0.306	0.551
ASIST (Microvilli-1)	1	50	0.586	0.311	0.556
ASIST (Microvilli-5)	5	50	0.660	**0.338**	0.627
ASIST (Microvilli-20)	20	50	**0.715**	0.332	**0.674**

T.V. is the number of training videos. T.F. is the number of training frames of each video. RSHN (Self) uses testing video for training. RSHN (Real) is the standard testing accuracy of using another independent video as training data.

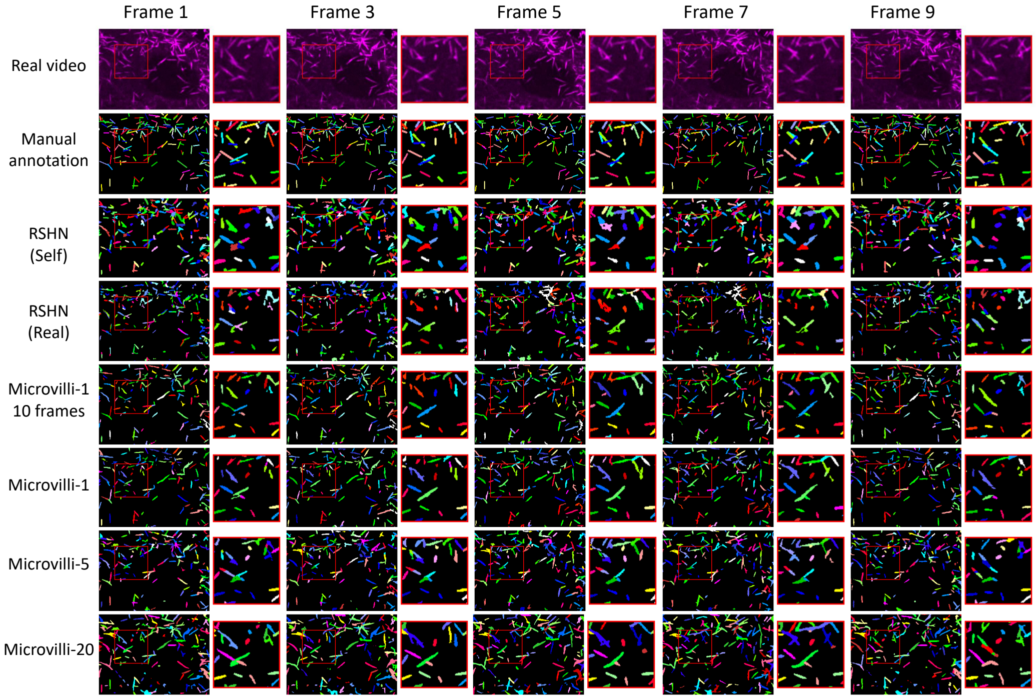


Figure 2.12: This figure shows the instance segmentation and tracking results of the real testing microvilli video.

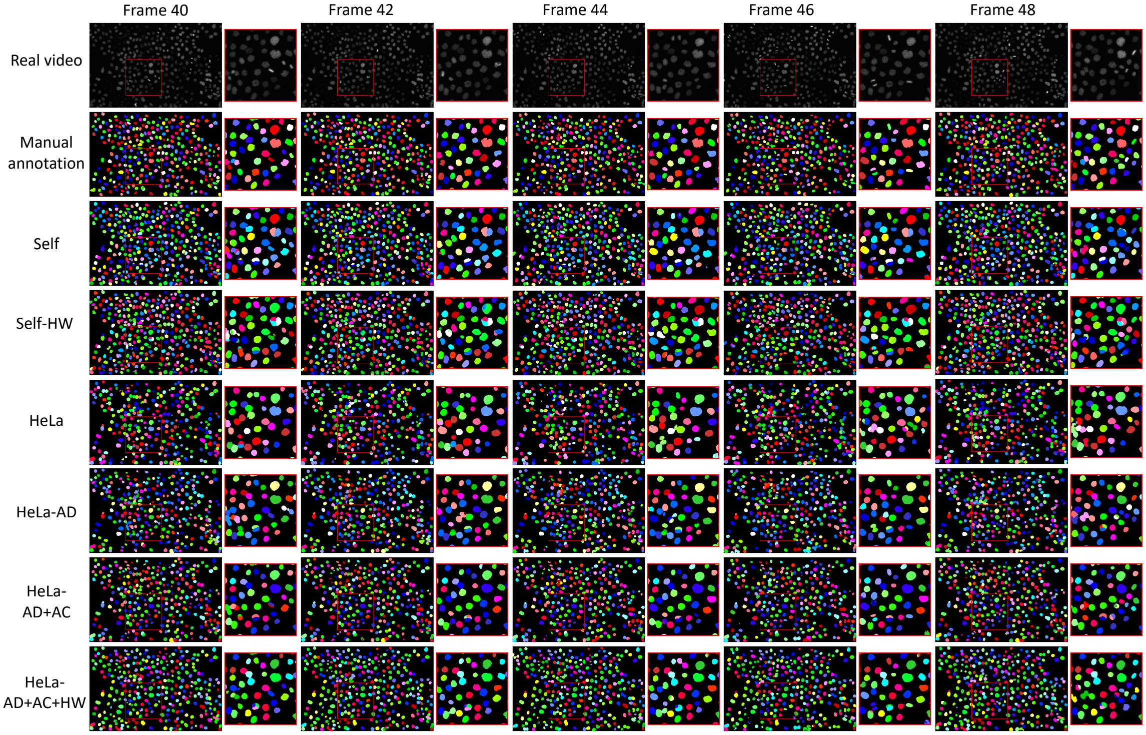


Figure 2.13: This figure shows the instance segmentation and tracking results on the real HeLa cell testing video.

2.2.5.2 Instance segmentation and tracking on HeLa cell videos

Instance segmentation and tracking results of HeLa cell videos were presented in Fig. 2.13. Based on the performance in Table. 2.3. HeLa-AD+AC+HW achieved superior performance than other settings using the ASIST method. The best performance of our annotation-free ASIST method is 5% to 9% lower than the manual annotation baseline. The most salient feature of improving the performance is to introduce the annotation cleaning (AC) step.

2.2.6 Discussion

In this paper, we assess the feasibility of performing pixel-embedding based instance object segmentation and tracking in an annotation-free manner, with adversarial simulations. Compared with conventional segmentation and tracking methods on microscope videos, our experiment used a pixel-embedding strategy instead of the "segmentation and association" two-step method. Our method also used synthetic training data instead of manual annotation. According to our experimental results, our annotation-free instance segmentation and tracking model achieved superior performance

Table 2.3: DET, SET and TRA values of different experiments on HeLa cell video.

Exp.	T.V.	T.F.	DET	SEG	TRA
RSHN (Self)	2	92	**0.979**	**0.884**	**0.975**
RSHN (Self-HW)	2	92	0.956	0.809	0.951
ASIST (HeLa)	10	50	0.858	0.656	0.849
ASIST (HeLa-AD)	10	50	0.853	0.718	0.844
ASIST (HeLa-AD+AC)	10	50	0.919	0.755	0.911
ASIST (HeLa-AD+AC+HW)	10	50	0.939	0.796	0.928

T.V. is the number of training videos. T.F. is the number of training frames per video. RSHN (Self) is the upper bound of RSHN using testing video for training.

on the microvilli dataset as well as comparable results on the HeLa dataset. Such encouraging results elucidated a promising new path to leverage the currently unsalable human annotation based pixel-embedding deep learning approach in an annotation free manner. In terms of robustness, the proposed pixel-embedding based method does not require heavy parameter tuning, which is typically inevitable in traditional model based methods. As a learning based method, the robustness of the proposed method can be further improved with more heterogeneous training images.

Strengths. The strength of our proposed ASIST method is three-fold: I. the proposed method is annotation-free to alleviate the extensive manual efforts of preparing large-scale manual annotations for training deep learning approaches; II. The proposed method does not require heavy parameter tuning; III. The proposed ASIST method combines the strength of both adversarial learning and pixel embedding based cell instance segmentation and tracking.

Limitations. One major limitation of our ASIST method is that both microvilli and HeLa cells have relatively homogeneous shape and appearance variations. In the future, it will be valuable to explore more complicated cell lines and more heterogeneous microscope videos. Meanwhile, the registration based method is introduced to capture the shape variations for ball-shaped HeLa cells. For more complicated cellular and subcellular objects, deep learning based solutions might be needed, such as the shape auto-encoder.

Following the proposed ASIST framework, our long term goal is to propose more general and comprehensive algorithms that can be applied to a variety of microscope videos with pixel-level instance segmentation and tracking. This would provide new analytical tools for domain experts to characterize high spatio-temporal dynamics of cells and subcellular structures.

2.2.7 Conclusion

In this paper, we propose the ASIST method – an annotation-free instance segmentation and tracking solution to characterize cellular and subcellular dynamics in microscope videos. Our method consists of unsupervised image-annotation synthesis, video synthesis, and instance segmentation and tracking. According to the experiments on subcellular (microvilli) videos and cellular (HeLa cell) videos, ASIST achieved comparable performance to manual annotation-based strategies. The proposed approach is a novel step towards annotation-free quantification of cellular and subcellular dynamics for microscope biology.

CHAPTER 3

Efficient Feature Representation Learning for Medical Optical Imaging

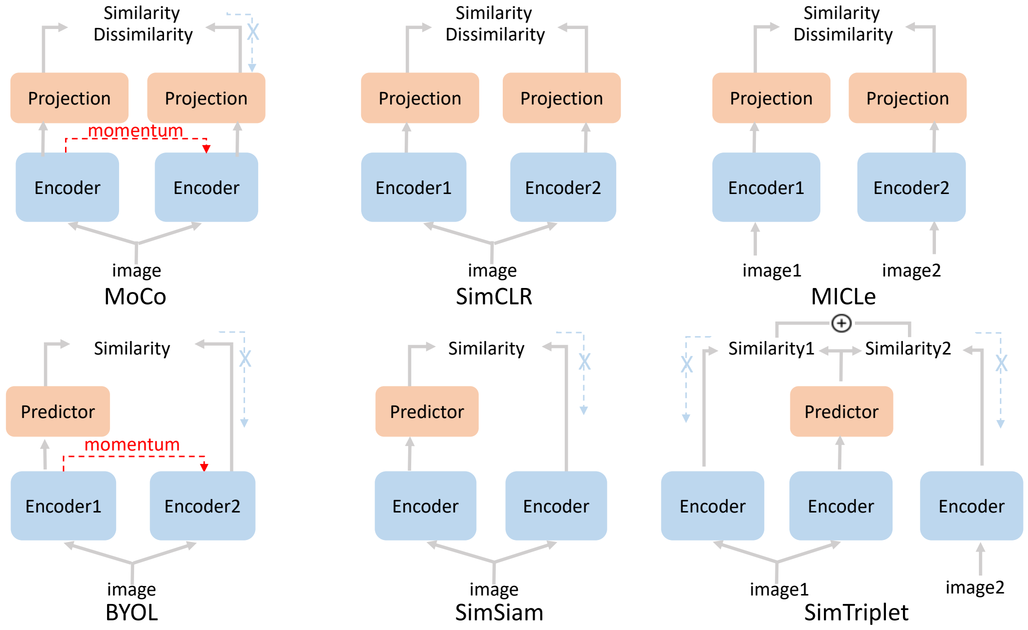
3.1 Simple Triplet Representation Learning on Histopathology Image with a Single GPU

3.1.1 Introduction

To extract clinically relevant information from GigaPixel histopathology images is essential in computer-assisted digital pathology (Liskowski and Krawiec, 2016; Zhu et al., 2017a; Xu et al., 2015b). For instance, the Convolutional Neural Network (CNN) based method has been applied to depreciate sub-tissue types on whole slide images (WSI) so as to alleviate tedious manual efforts for pathologists (Xu et al., 2017). However, pixel-wise annotations are resource extensive given the high resolution of the pathological images. Thus, the fully supervised learning schemes might not be scalable for large-scale studies. To minimize the need of annotation, a well-accepted learning strategy is to first learn local image features through unsupervised feature learning, and then aggregate the features with multi-instance learning or supervised learning (Hou et al., 2016a).

Recently, a new family of unsupervised representation learning, called contrastive learning (Fig. 3.1), shows its superior performance in various vision tasks (Wu et al., 2018; Noroozi and Favaro, 2016; Zhuang et al., 2019; Hjelm et al., 2018). Learning from large-scale unlabeled data, contrastive learning can learn discriminative features for downstream tasks. SimCLR (Chen et al., 2020b) maximizes the similarity between images in the same category and repels representation of different category images. Wu et al. (Wu et al., 2018) uses an offline dictionary to store all data representation and randomly select training data to maximize negative pairs. MoCo (He et al., 2020a) introduces a momentum design to maintain a negative sample pool instead of an offline dictionary. Such works demand large batch size to include sufficient negative samples (Fig. 3.1). To eliminate the needs of negative samples, BYOL (Grill et al., 2020) was proposed to train a model with a asynchronous momentum encoder. Recently, SimSiam (Chen and He, 2020) was proposed to further eliminate the momentum encoder in BYOL, allowing less GPU memory consumption.

To define different image patches as negative samples on pathological images is tricky since such a definition can depends on the patch size, rather than semantic differences. Therefore, it would be more proper to use nearby image patches as multi-view samples (or called positive samples) of the



	MoCo	SimCLR	MICLe	BYOL	SimSiam	SimTriplet (ours)
No negative samples	✗	✗	✗	✓	✓	✓
Multiple views	✗	✗	✓	✗	✗	✓
No momentum	✗	✓	✓	✗	✓	✓
Computing resource (from original papers)	8~64 32GB GPUs	32~128 TPU cores	16~64 TPU cores	64 TPU v3 cores	1~8 GPUs	1 GPU

Figure 3.1: Comparison of contrastive learning strategies. The upper panel compares the proposed SimTriplet with current representative contrastive learning strategies. The lower panel compares different approaches via a table.

same tissue type (Tian et al., 2019) rather than negative pairs. MICLe (Azizi et al., 2021) applied multi-view contrastive learning to medical image analysis. Note that in (Tian et al., 2019; Azizi et al., 2021), the negative pairs are still needed within the SimCLR framework.

In this paper, we propose a simple triplet based representation learning approach (SimTriplet), taking advantage of the multi-view nature of pathological images, with effective learning by using only a single GPU with 16GB memory. We present a triplet similarity loss to maximize the similarity between two augmentation views of same image and between adjacent image patches. The contribution of this paper is three-fold:

- The proposed SimTriplet method takes advantage of the multi-view nature of medical images beyond self-augmentation.
- This method minimizes both intra-sample and inter-sample similarities from positive image

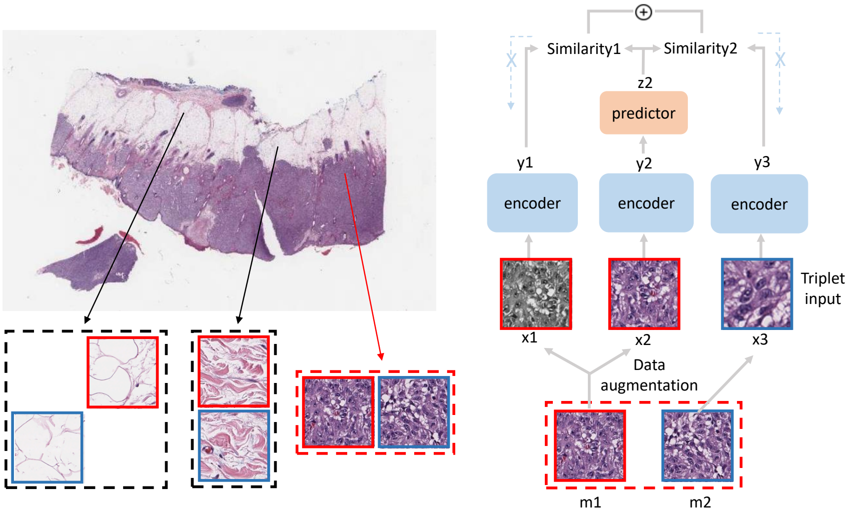


Figure 3.2: Network structure of the proposed SimTriplet. Adjacent image pairs are sampled from unlabeled pathological images (left panel) for triplet representation learning (right panel). The GigaPixel pathological images provide large-scale ”positive pairs” from nearby image patches for SimTriplet. Each triplet consists of two augmentation views from m_1 and one augmentation view from m_2. The final loss maximizes both inter-sample and intra-sample similarity as a representation learning.

pairs, without the needs of negative samples.

- The proposed method can be trained using a single GPU setting with 16GB memory, with batch size = 128 for 224×224 images, via mixed precision training.

3.1.2 Methods

The principle network of SimTriplet is presented in Fig 3.2. The original SimSiam network can be interpreted as an iterative process of two steps: (1) unsupervised clustering and (2) feature updates based on clustering (similar to K-means or EM algorithms) (Chen and He, 2020). By knowing the pairwise information of nearby samples, the SimTriplet aims to further minimize the distance between the ”positive pairs” (images from the same classes) in the embedding space (Fig. 3.3). In the single GPU setting with batch size 128, SimTriplet provides more rich information for the clustering stage.

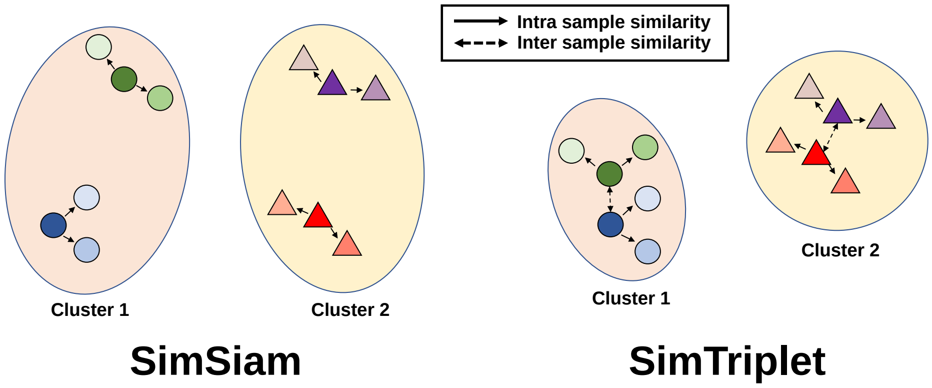


Figure 3.3: Compare SimTriplet with SimSiam. SimSiam network maximizes intra-sample similarity by minimizing the distance between two augmentation views from the same image. The proposed SimTriplet model further enforces the inter-sample similarity from positive sample pairs.

3.1.2.1 Multi-view nature of medical images

In many medical image analysis tasks, multi-view (or called multi-instance) imaging samples from the same patient or the same tissue can provide complementary representation information. For pathological images, the nearby image patches are more likely belong to the same tissue type. Thus, the spatial neighbourhood on a WSI provide rich ”positive pairs” (patches with same tissue types) for triplet representation learning. Different from (Hoffer and Ailon, 2015), all samples in our triplets are positive samples, inspired by (Chen and He, 2020). To train SimTriplet, we randomly sample image patches as well as their adjacent patches (from one of eight nearby locations randomly) as positive sample pairs from the same tissue type.

3.1.2.2 Triplet representation learning

Our SimTriplet network forms a triplet from three randomly augmented views by sampling positive image pairs (Fig. 3.2). The three augmented views are fed into the encoder network. The encoder network consists of a backbone network (ResNet-50 (He et al., 2016)) and a three-layer multilayer perceptron (MLP) projection header. The three forward encoding streams share the same parameters. Next, an MLP predictor is used in the middle path. The predictor processes the encoder output from one image view to match with the encoder output of two other image views. We applies stop-gradient operations to two side paths. When computing loss between predictor output and image representation from encoder output, encoded representation is regarded as constant (Chen and He, 2020). Two encoders on side paths will not be updated by back propagation. We used negative cosine similarity Eq.(3.1) between different augmentation views of (1) the same image patches, and (2) adjacent image patches as our loss function. For example, image m_1 and image m_2 are two adjacent patches cropped from the original whole slide image (WSI). x_1 and x_2 are randomly augmented views of image m_1, while x_3 is the augmented view of image m_2. Representation y_1, y_2 and y_3 are encoded from augmented views by encoder. z_1, z_2 and z_3 are the representation processed by the predictor.

$$\mathscr{C}(p,q) = -\frac{p}{\|p\|_2} \cdot \frac{q}{\|q\|_2} \tag{3.1}$$

$\mathscr{L}_{Intrasample}$ is the loss function to measure the similarities between two augmentation views x_1 and x_2 of image m_1 as seen in Eq.(3.2).

$$\mathscr{L}_{Intrasample} = \frac{1}{2}\mathscr{C}(y_1, z_2) + \frac{1}{2}\mathscr{C}(y_2, z_1) \tag{3.2}$$

$\mathscr{L}_{Intersample}$ is the loss function to measure the similarities between two augmentation views x_2 and x_3 of adjacent image pair m_1 and m_2 as in Eq.(3.3).

$$\mathscr{L}_{Intersample} = \frac{1}{2}\mathscr{C}(y_2, z_3) + \frac{1}{2}\mathscr{C}(y_3, z_2) \tag{3.3}$$

The triplet loss function as used in our SimTriplet network is defined as:

$$\mathscr{L}_{total} = \mathscr{L}_{Intrasample} + \mathscr{L}_{Intersample} \tag{3.4}$$

$\mathscr{L}_{Intrasample}$ minimizes the distance between different augmentations from the same image. $\mathscr{L}_{Intersample}$ minimizes the difference between nearby image patches.

3.1.2.3 Expand batch size via mix precision training

Mix precision training (Micikevicius et al., 2018) was invented to offer significant computational speedup and less GPU memory consumption by performing operations in a half-precision format. The minimal information is stored in single-precision to retain the critical parts of the training. By implementing the mix precision to SimTriplet, we can extend the batch size from 64 to 128 to train images with 224×224 pixels, using a single GPU with 16GB memory. The batch size 128 is regarded as a decent batch size in SimSiam (Chen and He, 2020).

3.1.3 Data and Experiments

3.1.3.1 Data

Annotated data. We extracted image patches from seven melanoma skin cancer Whole Slide Images (WSIs) from the Cancer Genome Atlas (TCGA) Datasets. From the seven annotated WSIs, 4698 images from 5 WSIs were obtained for training and validation, while 1,921 images from 2 WSIs were used for testing. Eight tissue types were annotated as: blood vessel (353 train 154 test), epidermis (764 train 429 test), fat (403 train 137 test), immune cell (168 train 112 test), nerve (171 train 0 test), stroma (865 train 265 test), tumor (1,083 train 440 test) and ulceration (341 train 184 test).

Following (Raju et al., 2020; Zhao et al., 2020b)), each image was a 512×512 patch extracted from 40× magnification of a WSI with original pixel resolution 0.25-micron meter. The cropped image samples were annotated by a board-certified dermatologist and confirmed by another pathologist. Then, the image patches were resized to 128×128 pixels. Note that the 224×224 image resolution provided 1.8% higher balance accuracy (based on our experiments) using supervised learning. We chose 128×128 resolution for all experiments for a faster training speed.

Unlabeled data. Beyond the 7 annotated WSIs, additional 79 WSIs without annotations were used for training contrastive learning models. The 79 WSIs were all available and usable melanoma cases from TCGA. The number and size of image patches used for different contrastive learning strategies are described in §**Experiment**.

3.1.3.2 Supervised learning

We used ResNet-50 as the backbone in supervised training, where the optimizer is Stochastic Gradient Descent (SGD) (Bottou, 2010) with the base learning rate $lr = 0.05$. The optimizer learning rate followed (linear scaling (Goyal et al., 2017)) $lr \times \text{BatchSize}/256$. We used 5-fold cross-validation by using images from four WSIs for training and images from the remaining WSI for validation. We trained 100 epochs and selected the best model based on validation. When applying the trained model to testing images, the predicted probabilities from five models were averaged. Then, the class with the largest ensemble probability was used as the predicted label.

3.1.3.3 Training contrastive learning benchmarks

We used the SimSiam network (Chen et al., 2020b) as the baseline method of contrastive learning. Two random augmentations from the same image were used as training data. In all of our self-supervised pre-training, images for model training were resized to 128 × 128 pixels. We used momentum SGD as the optimizer. Weight decay was set to 0.0001. Base learning rate was $lr = 0.05$ and batch size equals 128. Learning rate was $lr \times \text{BatchSize}/256$, which followed a cosine decay schedule (Loshchilov and Hutter, 2017). Experiments were achieved only on a single GPU with 16GB memory. Models were pre-trained for $39,500/128 \times 400 \approx 127,438$ iterations. 79 unlabeled WSIs were used for self-supervised pre-training. We randomly cropped 500 images from each WSI and resized them to 128 × 128 pixels. 39,500 images in total serve as the original training data.

Following MICLe (Azizi et al., 2021), we employed multi-view images as two inputs of the network. Since we did not use negative samples, multi-view images was trained by SimSiam network instead of SimCLR. For each image in the original training dataset, we cropped one patch which is randomly selected from its eight adjacent patches consisting of an adjacent images pairs. We had 79,000 images (39,500 adjacent pairs) as training data. Different from original SimSiam, network inputs were augmentation views of an adjacent pair. Referring to (Chen and He, 2020), we applied our data on SimSiam network. First, we used 39,500 images in original training dataset to pre-train on SimSiam. To see the impact of training dataset size, we randomly cropped another 39,500 images from 79 WSIs for training on a larger dataset of 79,000 images. We then used training data from the MICLe experiment to train the SimSiam network.

3.1.3.4 Training the proposed SimTriplet

The same 79,000 images (39,500 adjacent pairs) were used to train the SimTriplet. Three augmentation views from each adjacent pair were used as network inputs. Two augmentation views were from one image, while the other augmentation view was augmented from adjacent images. Three augmentation views were generated randomly, where the augmentation settings were similar with the experiment on SimSiam (Chen et al., 2020b). Batch size was 128 and experiment run on a single 16GB memory GPU.

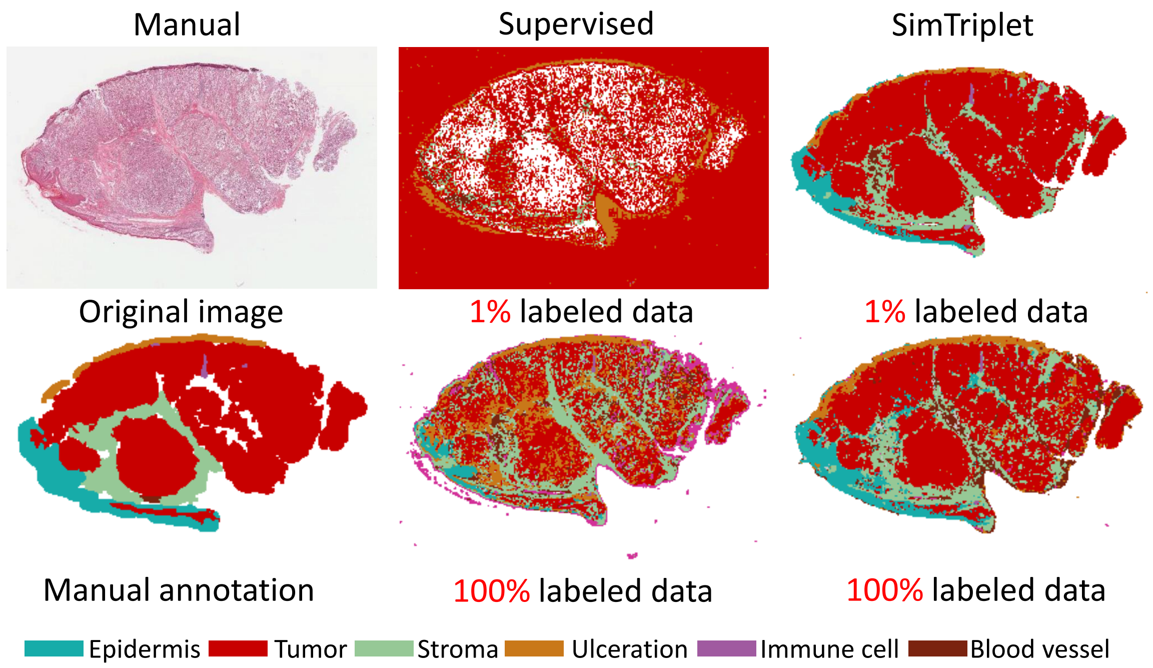


Figure 3.4: Visualization of classification results. One tissue sample is manually segmented by our dermatologist (via QuPath software) to visually compare the classification results. The contrasting learning achieved superior performance compared with supervised learning, even using only 1% of all available labeled data.

3.1.3.5 Linear evaluation (fine tuning)

To apply the self-supervised pre-training networks, as a common practice, we froze the pretrained ResNet-50 model by adding one extra linear layer which followed the global average pooling layer. When finetuning with the annotated data, only the extra linear layer was trained. We used the SGD optimizer to train linear classifier with a based (initial) learning rate lr=30, weight decay=0, momentum=0.9, and batch size=64 (follows (Chen and He, 2020)). The same annotated dataset were used to finetune the contrastive learning models as well as to train supervised learning. Briefly, 4,968 images from 5 annotated WSIs were divided into 5 folders. We used 5-fold cross validation: using four of five folders as training data and the other folder as validation. We trained linear classifiers for 30 epochs and selected the best model based on the validation set. The pretrained models were applied to the testing dataset (1,921 images from two WSIs). As a multi-class setting, macro-level average F1 score was used (Attia et al., 2018). Balanced accuracy was also broadly used to show the model performance on unbalanced data (Brodersen et al., 2010).

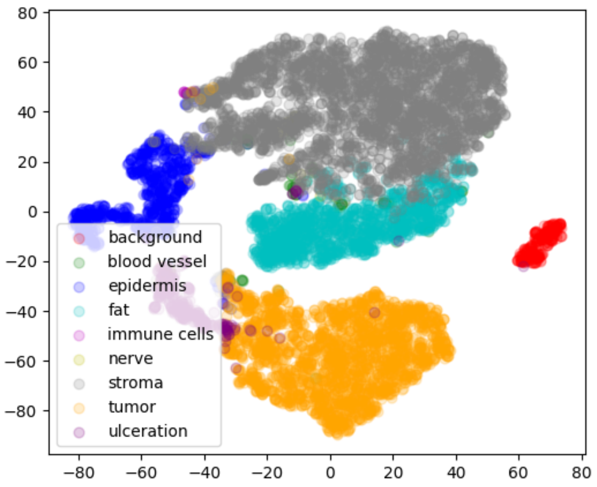


Figure 3.5: t-SNE plot of abstracted feature by SimTriplet model. The abstracted feature is shown in t-SNE plot. Different color dots represent different tissue types.

3.1.4 Results

3.1.4.1 Model classification performance.

F1 score and balanced accuracy were used to evaluate different methods as described above. We trained a supervised learning models as the baseline. From Table 3.3, our proposed SimTriplet network achieved the best performance compared with the supervised model and SimSiam network (Chen and He, 2020) with same number of iterations. Compared with another benchmark SwAV(Caron et al., 2020), the F1 score and balanced accuracy of SwAV are 0.53 and 0.60, which are inferior compared with our SimTriplet (0.65 and 0.72) using the same batch size = 128 within 16GB GPU memory. To show a qualitative result, a segmentation of a WSI from test dataset is shown in Fig. 3.4.

3.1.4.2 Model performance on partial training data.

To evaluate the impact of training data number, we trained a supervised model and fine-tuned a classifier of the contrastive learning model on different percentages of annotated training data (Table 3.2). Note that for 1% to 25%, we ensure different classes contribute a similar numbers images to address the issue that the annotation is highly imbalanced.

Table 3.1: Classification performance.

Methods	Unlabeled Images	Paired Inputs	F1 Score	Balanced Acc
Supervised	0		0.5146	0.6113
MICLe (Azizi et al., 2021)*	79k	✓	0.5856	0.6666
SimSiam (Chen and He, 2020)	39.5k		0.5421	0.5735
SimSiam (Chen and He, 2020)	79k	✓	0.6267	0.6988
SimSiam (Chen and He, 2020)	79k		0.6275	0.6958
SimTriplet (ours)	79k	✓	**0.6477**	**0.7171**

* We replace SimCLR with SimSiam.

Table 3.2: Balanced Acc of using different percentage of annotated data.

Methods	Percentage of Used Annotated Training Data			
	1%	10%	25%	100%
Supervised	0.0614	0.3561	0.4895	0.6113
SimSiam (Chen and He, 2020)	0.7085	0.6864	0.6986	0.6958
SimTriplet	**0.7090**	**0.7110**	**0.7280**	**0.7171**

3.1.5 Conclusion

In this paper, we proposed a simple contrastive representation learning approach, named SimTriplet, advanced by the multi-view nature of medical images. Our proposed contrastive learning methods maximize the similarity between both self augmentation views and pairwise image views from triplets. Moreover, our model can be efficiently trained on a single GPU with 16 GB memory. The performance of different learning schemes are evaluated on WSIs, with large-scale unlabeled samples. The proposed SimTriplet achieved superior performance compared with benchmarks, including supervised learning baseline and SimSiam method. The contrastive learning strategies showed strong generalizability by achieving decent performance by only using 1% labeled data.

3.2 Integrate Memory Efficiency Methods for Self-supervised Learning on Pathological Image Analysis

3.2.1 Introduction

As the practice of using a larger batch size for model training increases, the limited computing resources become the primary barrier to deep learning development. According to the developed neural network model, larger models with more parameters normally contribute to a better performance. The proposed wide ResNet (Zagoruyko and Komodakis, 2016) has better performance over ResNet (He et al., 2016) on ImageNet with larger model size and more parameters. In Natural Language Process field, BERT-Large (Devlin et al., 2018) model has a higher GLUE score than BERT-Base from 79.6 to 82.1 with approximately three times the parameters (from 110M to 340M). However, the grow speed of GPU memory size is not comparable with deep learning model size increment.

In terms of large-scale image analysis, models which take large-scale images as input need large GPU memory for model training due to two aspects. On the one hand, images for training are large-scale in multiple fields (e.g., pathology image and satellite images). Pathology images in the TCGA dataset (Tomczak et al., 2015b) have gigapixels per image. It is not adaptable to feed large scale data into the model directly. On the other hand, models that achieve state-of-the-art performance require sufficient computing resources. For example, within contrastive learning, the primary limitation is that contrastive learning methods need a large batch size to learn the similarity and dissimilarity between samples within the same batch. SimCLR (Chen et al., 2020b) employed 128 TPU v3 cores to train a model with a batch size of 4096. MoCo (He et al., 2020a) also needs 8 32GB GPUs to enable a 1024 batch size. Different from these two methods, BYOL (Grill et al., 2020) is less sensitive to batch size but still trained with 64 TPU v3 cores with a batch size of up to 4096.

As mentioned before, limited computing resources are a barrier; thus, multiple research works have been proposed to adapt model training on limited GPU resources. As developed in (Pal et al., 2019a; Le et al., 2011), model training and data processing can be deployed on multiple GPU devices which enable large batch size and speed up the training process. GPU parallel computing requires supported GPU devices and parallel computing mechanism. To mitigate the available GPU device limitation, an increased effort has been made to develop memory efficient training strategies on GPU. Song Han et al. (Han et al., 2016) reduce model size by pruning and Huffman coding.

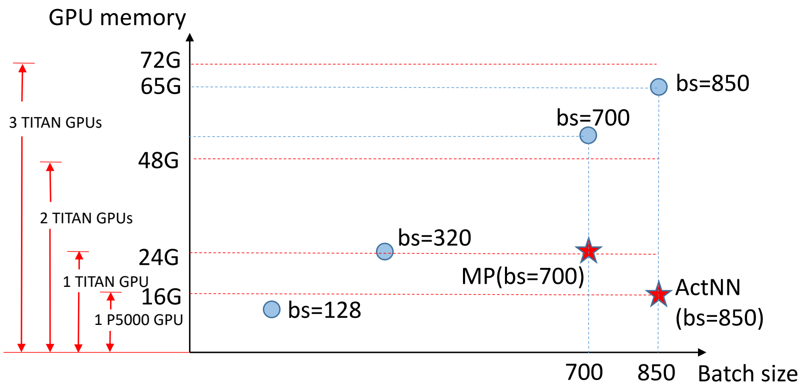


Figure 3.6: GPU maximizing efficiency comparison. Blue circles use no GPU memory-efficient methods. ”bs” means the batch size used in model training. MP is mixed precision training. Method on the lower right is preferred which achieves a larger batch size utilizing less GPU memory.

Yujun Lin et al. (Lin et al., 2017) proposed to compress the gradient in the training process in order to reduce the communication load in parallel training. NVIDIA also proposed mixed precision training (Micikevicius et al., 2017) to a half model weight and gradient precision in model training. Besides model size and gradient, ActNN (Chen et al., 2021a) is designed for activation value compression in training. A GPU maximizing efficiency comparison is shown in Fig.3.6.

For large-scale image analysis, both model performance, training speed, and GPU memory requirement are important. Current methods for maximizing memory efficiency normally choose to use low-precision data formats such as Mixed Precision Training (Micikevicius et al., 2017) which may harm the training accuracy. Extra operations on the gradient or the activation value will affect the training speed, such as in ActNN (Chen et al., 2021a). In this paper, we implement multiple memory-efficient training methods for pathology image analysis. We train the contrastive learning model BYOL (Grill et al., 2020) on a single GPU with memory-efficient methods and multiple GPUs with data parallel processing strategy. Meanwhile, we evaluate the model performance through the downstream classification tasks. The contribution of our study is three-fold: (1) We implement advanced memory efficient methods on self-supervised learning model. (2) We enabled contrastive learning on pathology images with limited computing resources (a single GPU). (3) We evaluate the performance on accuracy, GPU efficiency, and training speed across GPU efficiency maximizing methods.

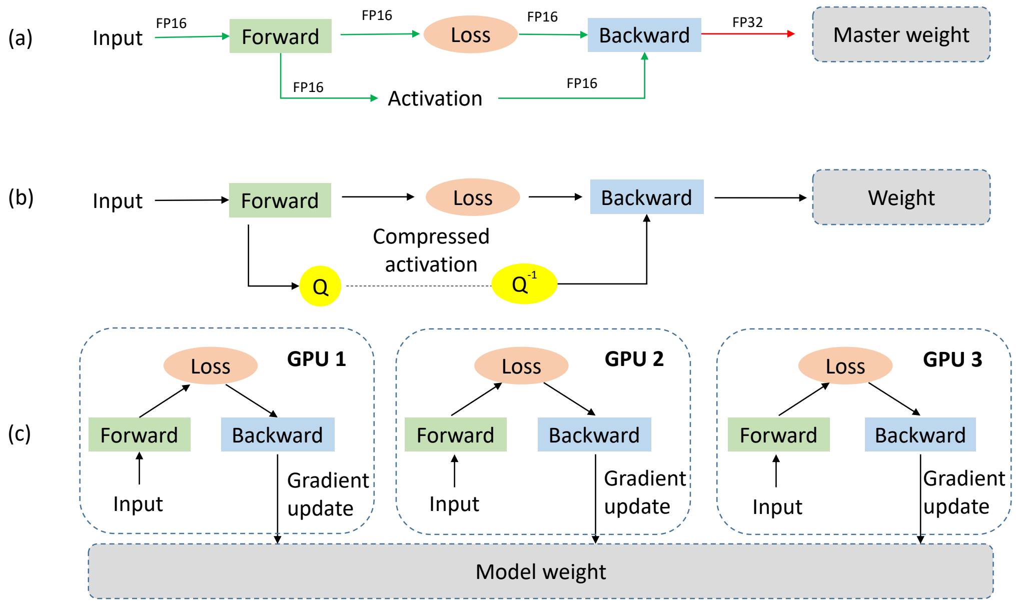


Figure 3.7: GPU maximizing efficiency methods pipelines. (a) Mixed precision training: In the forward and backward pass, FP16 is used for computing which halves bandwidth and GPU occupancy rate. (b) ActNN: Compress activations before storage and decompress activation for backpropagation. (c) Data parallel computing: Multiple GPU devices process data concurrently and update the shared model weights.

3.2.2 Method

To maximize GPU efficiency on contrastive learning, three model training settings are implemented: (1) mixed precision training, (2) ActNN, and (3) Data parallel computing. In large-scale image analysis, our goal is to classify multiple tissue patches without image annotation on a single GPU. Model pipelines of these methods are shown in Fig.4.11 .

3.2.2.1 BYOL

BYOL (Grill et al., 2020) is used as our contrastive learning model. As the contrastive learning designed purpose, BYOL is a self-supervised learn used to learn image representation from images without annotations. BYOL uses two neural networks (default network is ResNet-50), an online network and a target network to learn image representations. The online network is trained to predict the target network output based on different augmentation views from the same image. With a momentum mechanism, the target network parameter is updated based on the online network parameter by slow-moving average. To better learn representation from similar images, a larger batch size contributes to better model performance.

3.2.2.2 Data parallel computing

Data parallel is widely used as a parallel processing strategy on multiple devices. In popular deep learning frameworks such as Pytorch and Tensorflow, data parallel computing helps to the model and data in multiple GPU devices. Data in different GPU can be processed parallel. The model parameter will be shared across all available GPU devices through communication between networks. Data parallel computing enable large scale data processing or model training with a large batch size.

3.2.2.3 Mixed precision training

Mixed precision training introduces the usage of a half-precision floating point tensor in the model training process. Normally, deep learning model training uses a single-precision(FP32) format to store data. To save GPU memory, NVIDIA (Micikevicius et al., 2017) proposes using a half-precision(FP16) format in storing model weights and gradients. To prevent model performance loss with low precision data format, they also introduce techniques such as loss-scaling, in order to compensate for lower precision data effects.

3.2.2.4 ActNN

ActNN (Chen et al., 2021a) proposed to save randomly quantized activation for both gradient computation and model back propagation. In the model training process, the activation value of each layer is stored for further back propagation which takes up significant GPU memory. As designed by ActNN, the activation function is compressed before it is stored as a tensor and the decompress activation value is decompressed before it is used for back propagation. In the activation compressing process, ActNN quantizes FP32 to a 2-bit number which will not harm the model training convergence.

3.2.2.5 In-place Operation

In model training process, instead of saving a copy of tensors, metrics and activation, value will be directly changed by in-place operation. In-place operation helps to reduce GPU memory usage when operating on large amount of data.

3.2.3 Data and experimental design

3.2.3.1 Data

3.2.3.1.1 Annotated data

Image patches were extracted from seven skin cancer Whole Slide Images (WSIs) in the Cancer Genome Atlas (TCGA) Datasets. From all of the patches cropped from the WSIs, 4698 images belonging to 5 WSIs were used as training and validation sets. 1,921 images from the other 2 WSIs were used as a testing set. Eight tissue types in all image patches were annotated: immune cell (168 train, 112 test), stroma (865 train, 265 test), blood vessel (353 train, 154 test), nerve (171 train, 0 test), epidermis (764 train, 429 test), ulceration (341 train, 184 test), tumor (1,083 train 440 test) and fat (403 train, 137 test). According to (Raju et al., 2020), each image was extracted from 40× magnification of a WSI in size of 512×512 when the original pixel resolution is 0.25 micron meter. The image samples annotations were made by a board-certified dermatologist and confirmed by a pathologist. Note that all image patches were resized to 128×128 pixels in all experiments for a faster training speed.

3.2.3.1.2 Unlabeled data

Beyond the seven WSIs used in data annotation, another 79 WSIs were used for BYOL model training. We randomly cropped 1,000 image patches into the size of 128×128 from each WSI. 79,000 images were used as contrastive learning training data.

3.2.3.2 Experimental design

In the contrastive learning model training, we use BYOL model in a default setting (Grill et al., 2020). We use ResNet-50 as a neural network backbone and stochastic gradient descent (SGD) as an optimizer. To ensure the comparison fairness, we apply in-place operation on activation value when updating the backbone neural network. Base learning rate is set to 0.05 and momentum value is set to 0.9. We use NVIDIA TITAN RTX 24G in these experiments.

3.2.3.2.1 Oracle

With the baseline experiment used for the purpose of comparison, we implement BYOL on a single GPU and train the model with a relatively small batch size of 128. The experiment setting takes a

10G GPU memory. We applied no GPU-efficient method on oracle experiment.

3.2.3.2.2 In-place

Based on basic BYOL model, we used in-place setting in model activation value to save GPU memory usage. Due to GPU memory size limitation, we use 3 GPUs train model with large batch size. In-place activation setting is also applied to following GPU-efficiency methods.

3.2.3.2.3 Mixed precision training + In-place

To show the advantages of the GPU maximizing efficiency method, we implement mixed precision training on single GPU and achieved batch size of up to 700, which takes 23G GPU memory. For comparison, we trained BYOL on 3 of the same GPUs by data parallel training, with the same batch size of 700.

3.2.3.2.4 ActNN + In-place

Similarly, we evaluate ActNN's performance over data parallel computing. The ActNN method achieves a batch size of 850 with a 15G GPU memory. By comparison, the data parallel processing achieves a batch size of 850 with three GPUs, which takes 65G memory.

3.2.3.2.5 Linear evaluation (fine-tuning)

To apply the pre-trained ResNet to the downstream task (classification), as the common practice, we froze the ResNet-50 model and added one linear layer following the ResNet-50 output. In the finetuning process, only the linear layer was trained. We use SGD as the optimizer to train the linear classifier with a learning rate of 30, and a batch size of 64. When finetuning the linear classifier, we used a 5-fold cross validation method: 4 fold as training, and the other fold as validation. The linear layer is trained for 30 epochs and the best best linear classifier is selected according to validation performance. For the task of multiclass classification, we used the F1 score and balanced accuracy to evaluate the model performance.

3.2.4 Results

Considering different GPU-efficient strategies using single or three GPU devices, we performed six experiments by training six BYOL models (ResNet-50 backbone).

Table 3.3: Classification performance.

Methods	#GPU	Memory(G)	Batch Size	F1 Score	Balanced Acc
Oracle	1	10	128	0.54	0.55
In-place	1	24	320	0.57	0.58
In-place	3	52	700	**0.68**	0.67
In-place	3	65	850	0.59	0.66
In-place+Mixed Precision	1	24	700	0.57	0.63
In-place+ActNN	1	15	850	0.66	**0.72**

* Oracle means we use no GPU-efficiency method for large batch size training.

3.2.4.1 Model classification performance

To evaluate the model performance trained with different GPU setting, we test model on image classification after fine-tuning. F1 score and balanced accuracy of different training methods is shown in Table.3.3. We also show the GPU number used for model training and GPU memory usage.

From results are shown in Table.3.3, it is obvious the model with a larger batch size has better performance. ActNN with batch size 850 achieves the best balance accuracy performance of 0.72. In-place with batch size 700 achieves the best F1 score. Compared with methods without GPU-efficiency methods, mixed precision training, and ActNN enable model training with larger batch sizes on limited computing resources. For the model requiring large batch size (e.g., BYOL), GPU-efficiency methods achieve better model performance on a single GPU.

3.2.4.2 Model training speed

Another important factor in evaluating the method efficiency is the model training speed. The model training times for 400 epochs are shown in Fig.3.8. It is obvious that we find data parallel computing with a larger batch size has faster training speed. Mix precision training achieves a similar speed on a single GPU as compared with data parallel computing on three GPUs. In terms of training speed, ActNN is relatively slow when compared with other methods because of the activation compression and decompression operation.

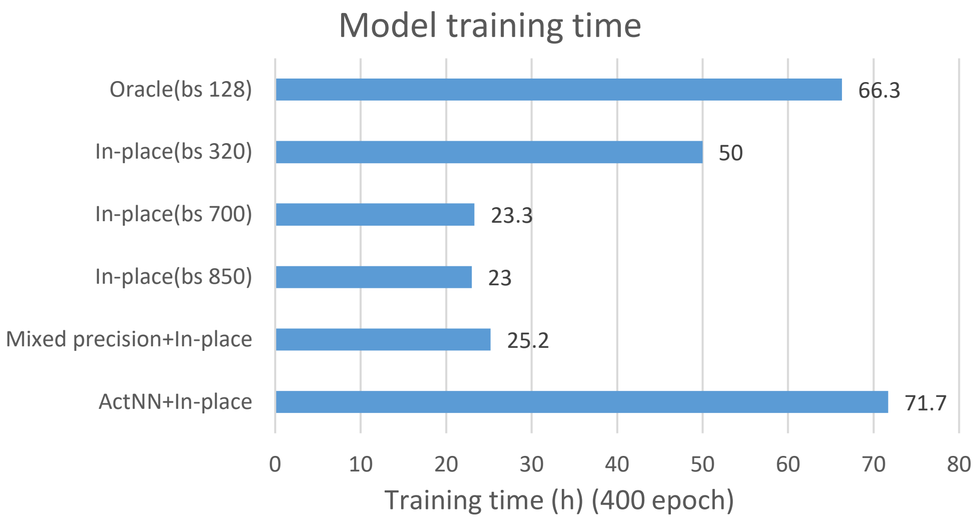


Figure 3.8: GPU maximizing efficiency methods training speed chart. ”bs” means training batch size used in model training. Oracle is implemented without GPU memory-efficient strategy and run parallel on multiple GPU devices.

3.2.5 Discussion

In this study, we combine GPU memory efficiency management strategy with self-supervised learning on large-scale image analysis. By deploying advanced GPU usage reduction methods, we achieve tripling training batch size which normally needs three GPU s parallel compering. From the experiments, implementing the GPU management strategy will not harm model performance. However, extra operations perform on the data will slow the training process down (e,g., ActNN). Training methods should take both training speed and model performance into consideration. The trade off is an essential point to discuss for GPU-efficiency methods in the future. GPU-efficient strategies can be integrated on other model training as well (e.g., transformer and graph neural network).

3.3 Leverage Supervised and Self-supervised Pretrain Models for Pathological Survival Analysis via a Simple and Low-cost Joint Representation Tuning

3.3.1 Introduction

Supervised pre-trained models (e.g., on ImageNet (Krizhevsky et al., 2012) and BiT (Lu et al., 2021c)) have been regarded as a powerful feature extractor and weight initializer in pathological image analysis (Chen et al., 2020a; Kieffer et al., 2017). However, it is resource-intensive to collect the large-scale annotated images, especially for gigapixel Whole Slide Images (WSIs) (Huo et al., 2021; David et al., 2019). Without requiring annotations, self-supervised learning (SSL) approaches are leading to a paradigm shift in large-scale pretraining for histopathological image analysis from visual inspection to more accurate quantitative assessment (Wang et al., 2021c; Yang et al., 2021; Liu et al., 2021a; Ciga et al., 2021), with the rapid growth of publicly available large-scale datasets (e.g., The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015b), and Pathology AI Platform (PAIP) (Kim et al., 2021)). In a recent study, Wang et al. (Wang et al., 2021c) utilized the entire TCGA and PAIP dataset to perform a self-supervised pretraining via a vision transformer, called TransPath. TransPath learned the pathological domain-specific information and achieved superior tissue classification performance.

However, most existing studies focused on developing more powerful pretrained models (Bao et al., 2021; Bardes et al., 2021; Tellez et al., 2018; Mormont et al., 2020), whose resource consumption is increasingly unscalable for academic institutes. Very few, if any, studies have investigated how to take advantage of existing, yet heterogeneous pretrained models for better performance on downstream tasks. As an example, the pathological data-optimized contrastive learning model TransPath (Wang et al., 2021c) achieved a superior performance compared with supervised models (e.g., ImageNet pretrained) on a classification cohort (Table 3.5). **Surprisingly, it yielded a inferior performance when it was translated to a cancer prognosis task.** Such phenomenon inspired us to explore how to leverage the already trained supervised and self-supervised models for pathological survival analysis (Zhu et al., 2016; Li et al., 2018; Tang et al., 2019).

In this paper, we propose a simple joint representation tuning (JRT) approach to aggregate task-agnostic vision representation (supervised ImageNet pretrained models) and pathological specific representation (self-supervised TCGA pretrained models) for downstream tasks (Fig. 3.9). This study also evaluated the different strategies as well as their performance of using heterogeneous

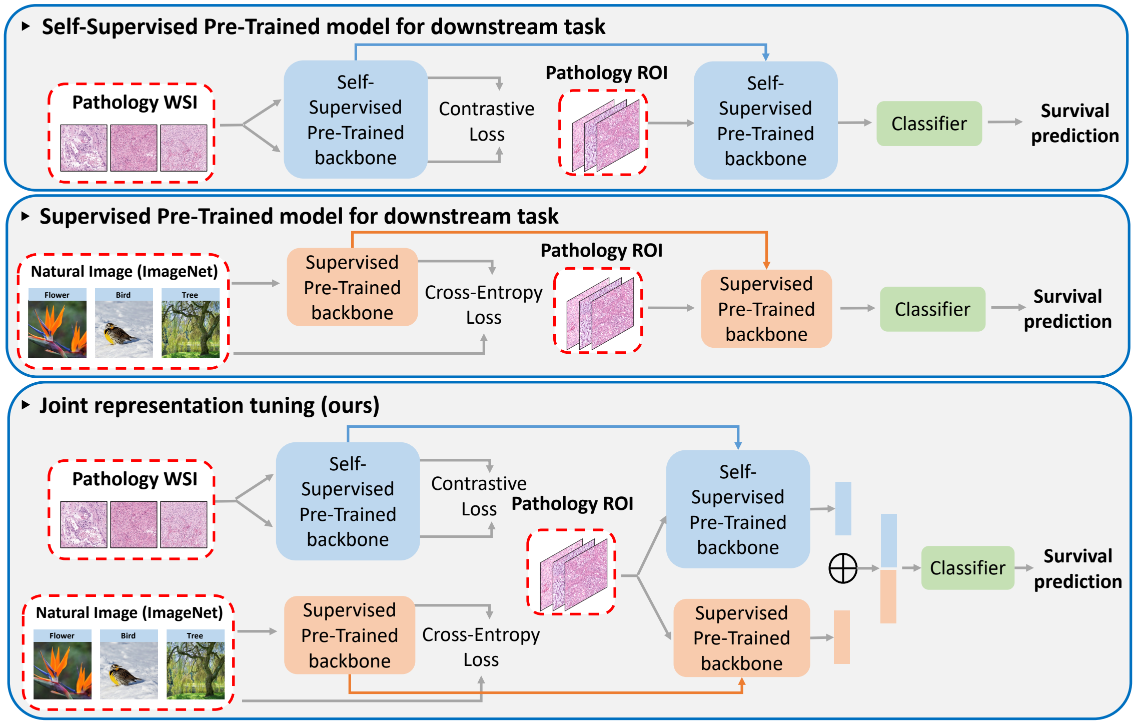

Self-Supervised Pre-Trained model for downstream task
Pathology WSI
Self-Supervised Pre-Trained backbone
Contrastive Loss
Pathology ROI
Self-Supervised Pre-Trained backbone
Classifier
Survival prediction
Supervised Pre-Trained model for downstream task
Natural Image (ImageNet)
Flower
Bird
Tree
Supervised Pre-Trained backbone
Cross-Entropy Loss
Pathology ROI
Supervised Pre-Trained backbone
Classifier
Survival prediction
Joint representation tuning (ours)
Pathology WSI
Self-Supervised Pre-Trained backbone
Contrastive Loss
Pathology ROI
Self-Supervised Pre-Trained backbone
Natural Image (ImageNet)
Flower
Bird
Tree
Supervised Pre-Trained backbone
Cross-Entropy Loss
Supervised Pre-Trained backbone
Classifier
Survival prediction

Figure 3.9: Model pipeline. In the top section, self-supervised model is pretrained with pathology WSIs and finetuned on pathology images for survival prediction. Middle section is supervised pretrained model with natural images of ImageNet and finetuned on pathology images. Our JRT method aggregates both pretrained models to achieve better downstream task performance.

pretrained models. The feature-direct JRT (f-JRT) that directly finetune the joint feature representations without 1) encoder network, 2) data augmentation, and 3) large memory consumption, achieved 60× speedup with decent performance on the survival analysis.

The contribution of this paper is in three-fold:

• The JRT method adapts and aggregates the task-agnostic vision representation (supervised ImageNet pretrained models) and pathological specific features presentation (self-supervised pretrained models) for better performance on downstream tasks.

• Comprehensive analyses on prevalent strategies of using heterogeneous pretrained models are conducted as a reference for the community.

• The joint representation tuning provides a simple, yet computationally efficient perspective to leverage large-scale pretained models for both cancer diagnosis and prognosis without extra resource-intensive pretraining.

3.3.2 Methods

The overall framework of the proposed JRT approach is presented in Fig. 3.9. We have also conducted a comprehensive analyses to evaluate a variety of (1) pretraining approaches, (2) feature extraction and finetuning strategies, (3) joint representation tuning methods, and (4) downstream tasks.

3.3.2.1 Supervised and self-supervised pretraining

Supervised and self-supervised pretraining are two prevalent vision representation learning strategies for downstream pathological image analysis (Thongprayoon et al., 2020; Peikari et al., 2018; Azizi et al., 2021). With a supervised learning procedures, ResNet (He et al., 2016) and VGG (Simonyan and Zisserman, 2015) can be pretrained by ImageNet, which have been widely used in medical image analysis (Rai et al., 2019; Bar et al., 2015). On the other hand, self-supervised learning (e.g., TransPath (Wang et al., 2021c)) has been increasingly popular for large-scale pathological pretraining. In our JRT method, we employed the ImageNet pretrained ResNet50 and TCGA+PAIP pretrained TransPath as encoders. We directly used the pretrained model weights for the downstream task.

3.3.2.2 Joint representation tuning

The framework of the proposed JRT is presented in Fig. 3.9 and 3.10. The low dimensional features from both supervised and self-supervised models are concatenated to a simple Multi-Layer Perception (MLP) for the downstream survival and diagnosis analyses (Jarrett et al., 2019). To evaluate the representation quality abstracted by a pretrained encoder, we conducted the survival prediction analysis as downstream task. The deep survival loss (Yao et al., 2020) was used as the loss function. We used Concordance Index (C-Index) (Uno et al., 2011) as an evaluation metric for our survival prediction. C-Index is defined as the ratio of the predicted survival time in correct order among all uncensored testing samples.

3.3.2.3 Evaluate different strategies of using pretrained model

We evaluated different ways of utilizing the pretrained models as baselines as shown in Fig. 3.10. To utilize the pretrained model, there are three basic strategies:

No-freeze. All weights in the pretrained network are freely finetuned using downstream task data. In this case, the pretrained weights are only used as weight initialization.

Encoder-freeze. The Encoder-freeze strategy freezes the encoder (e.g., convolutional encoder or vision transformer-based encoder) without further changing its weights. Then, the finetuning is only performed on features.

Feature-direct. Feature-direct strategy is designed as a ”two-stage” framework, where the feature encoders are discarded after extracting features. Then the extracted features are directly used for the downstream classification. The advantage of the Feature-direct strategy is that the memory consumption is minimized without using the encoders. However, this strategy is limited by not performing an on-line data augmentation.

3.3.3 Experiments

3.3.3.1 Data description

3.3.3.1.1 Survival prediction task

Two available public datasets were used in this study. The WSI images used for the survival prediction task were from TCGA-GBMLGG (Mobadersany et al., 2018). Each image was 1024×1024 resolution. 1505 ROI images from 769 patients were used for prognosis prediction model tuning. ROI patches were curated in (Mobadersany et al., 2018) from diagnostic slides. We randomly cropped 512×512 image patches from the ROI images.

3.3.3.1.2 Classification task

NCT-CRC-HE (Kather et al., 2019b) from National Center for Tumor Diseases (NCT) was used as the classification dataset (8 colorectal cancer types and normal) with 100,000 images. Images were in size of 224×224 from 86 WSIs.

3.3.3.2 Experimental setting

Our proposed joint finetuning strategy were designed to finetune the model with both CNN features and Transformer features. We used TransPath model (Wang et al., 2021c) and ResNet-50 (Chen et al., 2020a) as the backbones. To fairly evaluate the our joint representation finetuning method, We utilized the same MLP structure as the downstream network for all experiments. The MLP was

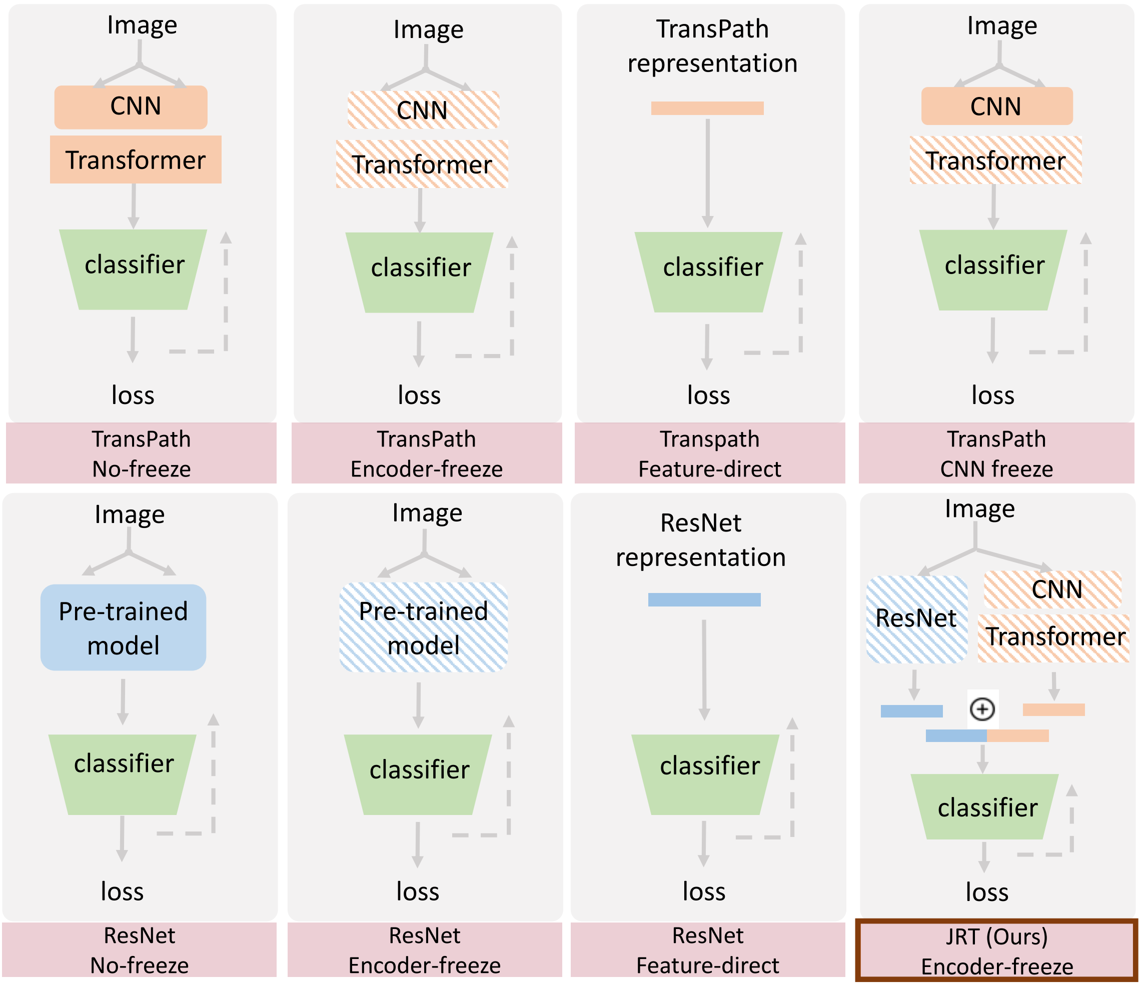


Figure 3.10: Finetuning strategies on pre-trained backbone. This figure shows the three prevalent finetuning strategies: (1) No-freeze, (2) Encoder-freeze, and (3) Feature-direct using ResNet and TranPath. The highlighted method is the proposed JRT strategy.

composed of three fully connected layers. As presented in Fig. 3.10, the image augmentation could be applied to No-freeze and Encoder-freeze training strategies for both TransPath and ResNet-50. We used Cox Loss (Yao et al., 2020) for model survival time prediction and Adam optimizer to update MLP. All model training were implemented on NVIDIA P5000 GPU.

3.3.4 Results

3.3.4.1 Performance on classification and survival prediction

The Table 3.4 and 3.5 indicated the performance of the proposed the JRT method as well as the benchmarks in survival prediction and tissue classification tasks. The results indicated that our JRT

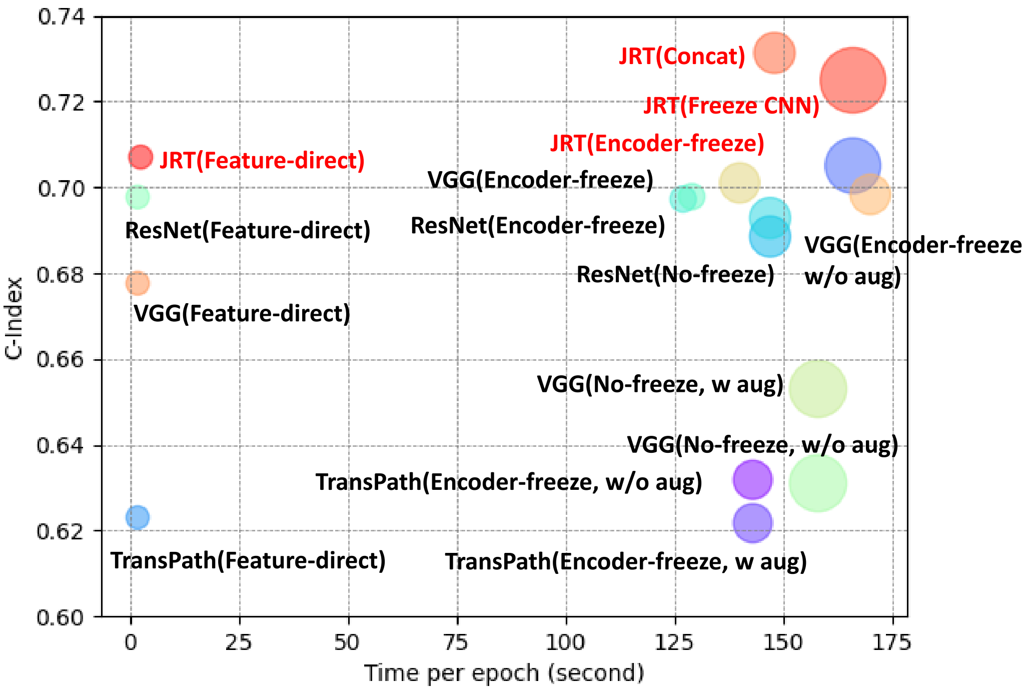


Figure 3.11: Visual comparison of different strategies. This figure shows the C-Index versus the model training time required for an epoch. The size of the blobs is proportional to the number of GPU memory consumption.

method achieved superior performance compared with the baseline methods aross two tasks.

The results indicated that the TransPath achieved the superior performance in the classification task compared with ResNet-50. However, it yielded the inferior performance in the survival prediction task compared with ResNet-50. The potential explanation would be the self-supervised pretraining was typically defined as a general classification task (classify if two augmented images were originally same). Therefore, the features might be over optimized for classification, while losing essential visual information for prognosis. By combining general vision features and pathological specific features for a better survival prediction and classification performance.

3.3.4.2 Computational resource

Fig. 3.11 presented training speed, C-index performance, and GPU memory consumption for different methods. Note that by using the proposed JRT method, the Feature-direct version (f-JRT) achieved more than 60× training speedup while maintaining 0.707 c-index score compared with

Table 3.4: Survival prediction performance on TCGA-GBMLGG ROI dataset.

Model	model freeze part	C-Index
VGG-16 (Simonyan and Zisserman, 2015)	Encoder-freeze	0.7010
ResNet-50 (He et al., 2016)	Encoder-freeze	0.6972
TransPath (Wang et al., 2021c)	Encoder-freeze	0.6217
JRT (Ours)	Encoder-freeze	**0.7313**

Table 3.5: Classification performance on NCT-CRC-HE dataset.

Method	F1-score	Accuracy
CNN (Wang et al., 2021c)*	0.9099	0.9081
ResNet-50 (He et al., 2016)	0.9541	0.9558
TransPath (Wang et al., 2021c)*	**0.9582**	0.9585
JRT (Ours)	0.9576	**0.9673**

* The experiment results are directly from (Wang et al., 2021c).

non-JRT methods.

3.3.5 Ablation Studies

The three strategies of using pretrained models were evaluated in the Table 3.6, including (1) No-freeze, (2) Encoder-free, and (3) Feature-freeze. Since the data augmentation could be applied or not for the No-freeze and Encoder-freeze approaches, we also evaluated the performance of adding data augmentation of downstream finetuning as ablation studies.

3.3.5.1 Different strategies of using pretrained models

From the performance on survival prediction analysis (Table 3.6), the Encoder-free strategy (with finetuning without updating backbone parameter) achieved similar performance with No-freeze strategy. Thus, the two-stage model yielded the similar performance as the more computational expensive end-to-end training. With two-stage design (Feature-direct), the three single backbone and our JRT method required less GPU memory and training time.

3.3.5.2 Effect of transformer

We compared the transformer based TransPath method with the CNN based ResNet-50 method. The results indicated that the TransPath achieved the superior performance in the classification task compared with ResNet-50. However, it yielded the inferior performance in the survival prediction task compared with ResNet-50.

3.3.5.3 Effect of data augmentation

To evaluate the effect of w/wo data augmentation, we applied random translation and rotation on ROI images. We compared such results with the ones without random translation applied. Experiments were implemented with TransPath, Resnet-50 and VGG-16 models, which were finetuned by brain cancer WSI ROI patches. The experiment with random translation applied achieved similar prognosis prediction. The result indicated that the data augmentation did not vary the performance in a noticeable margin. The explanation might be that the pretraining stage had already incorporated the corresponding data augmentation.

Table 3.6: Comparison of volumetric analysis metrics between the proposed method and the state-of-the-art clinical study on kidney components.

Backbone	Model freeze part	Augmentation	Memory(G)	Time(s)	C-Index
TransPath	No-freeze	w	7.05	172	N/A
TransPath	No-freeze	w/o	7.05	172	N/A
TransPath	Encoder-freeze	w	2.94	143	0.6318
TransPath	Encoder-freeze	w/o	2.94	143	0.6217
TransPath	CNN part	w/o	6.00	166	0.7050
TransPath	Feature-direct	w/o	1.03	1.74	0.6230
ResNet-50	No-freeze	w	3.28	147	0.6885
ResNet-50	No-freeze	w/o	3.28	147	0.6928
ResNet-50	Encoder-freeze	w	1.32	127	0.6972
ResNet-50	Encoder-freeze	w/o	1.32	129	0.6978
ResNet-50	Feature-direct	w/o	1.05	1.75	0.6977
VGG-16	No-freeze	w	6.32	158	0.6310
VGG-16	No-freeze	w/o	6.32	158	0.6530
VGG-16	Encoder-freeze	w	3.13	140	0.7010
VGG-16	Encoder-freeze	w/o	3.13	144	0.6983
VGG-16	Feature-direct	w/o	1.05	1.75	0.6776
f-JRT (Ours)	Feature-direct	w/o	1.07	2.44	0.7070
JRT (Ours)	TransPath CNN	w	8.331	166	0.7249
JRT (Ours)	Encoder-freeze	w	3.28	148	**0.7313**

”w” in augmentation column means training data are augmented image views. ”w/o” means all image views are fed into model without image view augmentation.

3.3.6 Conclusion

In this work, we analyze how to leverage the already trained supervised and self-supervised models for pathological survival analysis. We proposed a simple and low-cost JRT representation tuning strategy and shows effective improvement to adapt classification based supervised and self-supervised representation for survival prediction. With the proposed JRT, the Feature-direct fine-tuning strategy yields 60× training speedup while maintaining superior c-index score compared with non-JRT methods.

3.4 Bayesian-based Multimodal Multi-level Fusion on Colorectal Cancer Microsatellite Instability Prediction

3.4.1 Introduction

Microsatellite instability (MSI) in colorectal cancer (CRC) determines whether patients with cancer respond exceptionally well to immunotherapy (Sahin et al., 2019). Because universal MSI testing requires additional complex genetic or immunohistochemical tests, it is not possible for every patient to be tested for MSI in clinical practice. Therefore, a critical need exists for broadly accessible, cost-efficient tools to aid patient selection for testing.

Deep learning-based methods have been successfully applied for automated MSI prediction directly from hematoxylin and eosin (H&E)-stained whole-slide images (WSIs) (Kather et al., 2019a; Yamashita et al., 2021). Kather et al. (Kather et al., 2019a) developed ResNet-based model to predict patients with MSI and MSS tumors. Another work (Yamashita et al., 2021) further proposed MSINet and proved the deep learning model exceeded the performance of experienced gastrointestinal pathologists at predicting MSI on WSIs. Despite the vital role of such diagnostic biomarkers (Sidaway, 2020), patients with similar histology profiles can exhibit diverse outcomes and treatment responses. Novel and more specific biomarkers are needed from a whole spectrum of modalities, ranging from radiology (Pei et al., 2022b; Wu et al., 2019; Echle et al., 2021), histology (Ushizima et al., 2022; Kather and Calderaro, 2020; Wang et al., 2022a), and genomics (Lipkova et al., 2022; Braman et al., 2021).

Given the large complexity of medical data, there are new trends to integrate complementary information from diverse data sources for multimodal data fusion (Chen et al., 2022b; Feng et al., 2022; Cui et al., 2022). Many models have shown the use of radiology data to consider macroscopic factors could achieve more accurate and objective diagnostic and prognostic biomarkers for various cancer types (Wang et al., 2019; He et al., 2020b; Yao et al., 2023; Dong et al., 2020). However, when integrating radiology images and WSIs for predicting MSI, the large data heterogeneity gap between the two modalities exists and makes the integration very difficult. Specifically, a WSI consists of tens of thousands of patches (Chen et al., 2021c; Lu et al., 2021b; Wei et al., 2019) while radiology data usually form with 3D shape (Golia Pernicka et al., 2019). How to design an effective fusion strategy and learn important interactions between radiology and pathology images is important but still remains unknown for MSI prediction in CRC.

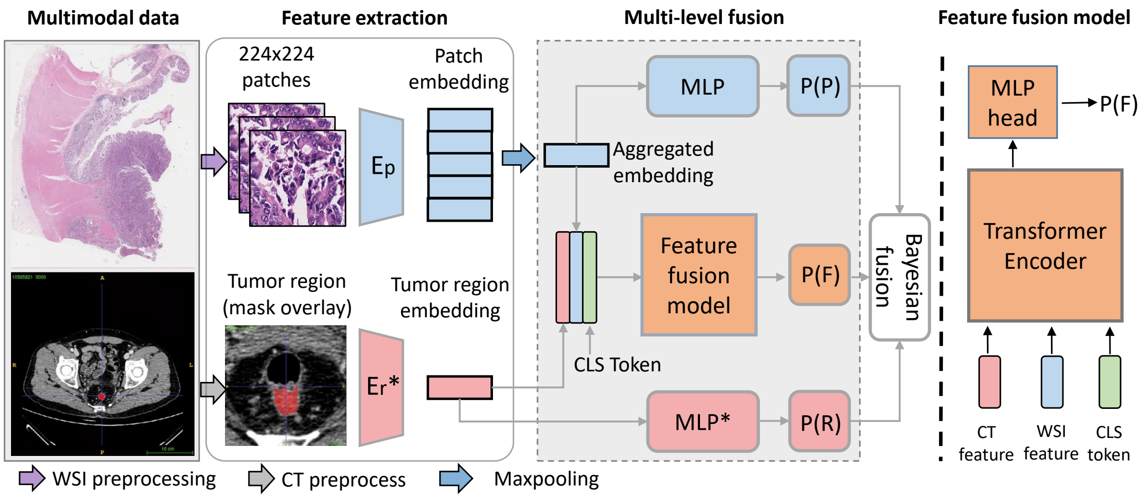


Figure 3.12: Our proposed M^2Fusion model. Multimodal data, WSI, and CT images are preprocessed to pathology image patches and CT tumor ROI, respectively. Embeddings are extracted by encoder E_p and E_r. $*$ means the model is well-trained and frozen in pipeline training. $\mathscr{P}_P$ is the pathology uni-model performance $\mathscr{P}(P_{ath})$. P_R is the radiology uni-model performance $\mathscr{P}(R_{ad})$. $\mathscr{P}_F$ is the feature level fusion model probability distribution under pathology and radiology guidance $\mathscr{P}(F_{ea}|P_{ath}R_{ad})$. The final fusion model by P_P, P_R and P_F is $\mathscr{P}(F_{ea}P_{ath}R_{ad})$ in Eq.3.8

In this paper, we introduce a new and effective multi-modal fusion pipeline for MSI prediction by combining decision-level fusion and feature-level fusion following Bayesian rules. We also investigated different fusion strategies and found the proposed fusion scheme achieved better results than those methods. The contributions of this paper are: 1) This study generalizes an MSI prediction pipeline in CRC utilizing radiology-guided knowledge. 2) To the best of our knowledge, we are the first to exploit a multi-level fusion strategy for using multi-modal data for MSI prediction. 3) Extensive experimental results suggest the effectiveness of our Bayesian-based multimodal multi-level fusion. It can reduce the gap between pathology and radiology predictions and achieve more robust and accurate fusions than other feature-level or decision-level methods.

3.4.2 Method

Problem Statement. In our study, each CRC patient has a 3D CT image, a pathology whole slide image (WSI), and its corresponding label (MSI status). We aim at CRC MSI prediction using both pathology and radiology data. Fig.3.12 shows the proposed Bayesian-based fusion model. Our fusion model combines three predictions together and can be seen as feature-level and decision-level fusion in a unified framework. It consists of two branches that process each modality (pathology or radiology data) and it introduces a radiology feature-guided pathology fusion model. In the following parts, we will discuss why radiology-guided fusion methods could benefit our final prediction.

3.4.2.1 Bayesian-based multi-modality fusion model

Assuming the learnable context from each modality is different, we hypothesize that the fusion between modalities knowledge can enhance the confidence level of the CRC MSI prediction, compared with single modality training. Due to the inherent scale difference between the two modalities (2D gigapixel WSI and 3D CT images), we propose a multi-modal fusion strategy, which combines both the decision-level prior and feature-level prior to enhance the interaction between the learnable knowledge from different fields of view.

We first define the predictions from pathology data and from radiology data as events P_{ath} and R_{ad}, respectively. Here, we hypothesize the probabilistic relationship between prediction with Bayes' theorem as follows:

$$\mathscr{P}(P_{ath}R_{ad}) = \mathscr{P}(R_{ad})\mathscr{P}(P_{ath}|R_{ad}) \tag{3.5}$$

Here $\mathscr{P}(R_{ad})$ is the uni-model performance on radiology data. $\mathscr{P}(P_{ath}|R_{ad})$ denotes the probabilistic prediction on the model well-trained on pathology data with radiology prior. According to Eq.3.5, if under the guidance of pre-trained radiology model $\mathscr{P}(R_{ad})$, pathology model $\mathscr{P}(P_{ath}|R_{ad})$ performs better than uni-model on pathology ($\mathscr{P}(P_{ath})$), then modality fusion model should perform better than uni-model ($\mathscr{P}(P_{ath})$ and $\mathscr{P}(R_{ad})$).

$$\mathscr{P}(P_{ath}R_{ad}) \propto \mathscr{P}(P_{ath}|R_{ad}) \tag{3.6}$$

The Bayes' theorem can be extended to three events: feature level multi-modal fusion model predicts MSI status correct as event F_{ea}. The extended Bayes' theorem is Eq.3.7.

$$\mathscr{P}(F_{ea}P_{ath}R_{ad}) = \mathscr{P}(F_{ea}|P_{ath}R_{ad})\mathscr{P}(P_{ath}R_{ad}) \tag{3.7}$$

Similar to the relation between $\mathscr{P}(P_{ath}|R_{ad})$ and $\mathscr{P}(P_{ath}R_{ad})$, Eq.3.8. If radiology data can help to get a better feature-level fusion model $\mathscr{P}(F_{ea}|P_{ath}R_{ad})$, the final fusion on both the decision-level and feature-level should outperform the decision-level fusion model.

$$\mathscr{P}(F_{ea}|P_{ath}R_{ad}) \propto \mathscr{P}(F_{ea}|P_{ath}R_{ad}) \tag{3.8}$$

Bayes' theorem guarantees that if we want to seek a better final fusion model than decision-level fusion, we have to implement a good feature-level fusion model. Our final model could benefit from both feature-level and decision-level fusion.

3.4.2.2 MSI prediction on single modality

Pathology model. Our pathology model is composed of two parts: First, we used the CLAM model(Lu et al., 2021b) to crop the pathology patches from gigapixel WSI. Second, following the previous work (Yamashita et al., 2021), the ResNet-18 is used as an encoder to abstract features from pathology patches. We crop the non-overlapping image tiles in size of 224×224 from the WSI foreground. The image patches from all WSI are constructed as a whole pathology patch dataset. The pathology patches label is inherited from the WSI label which it cropped from. The model will predict a patch-level probability of whether the patches belong to MSI or MSS. In the testing phase, the image patches will get the predicted label from the well-trained encoder. The majority vote result of patches from WSI is the patient MSI prediction.

Radiology model. Based on the 3D radiology CT scans, the tumor region mask of CT volume has been annotated. Two essential slices are cropped from three directions of CT image. One slice is CT tumor region by overlaying the mask on the CT slice. The other slice is the whole CT slice in the direction. The six essential slices (two slices from each direction) are stacked as a six-channel input to build a 2.5D model (Roth et al., 2014). The encoder used for MSI prediction is ImageNet pre-trained ResNet-18 (modified input channel to six channels). The original 3-channel pre-trained weights are copied to 4^{th} to 6^{th} channel as initialization.

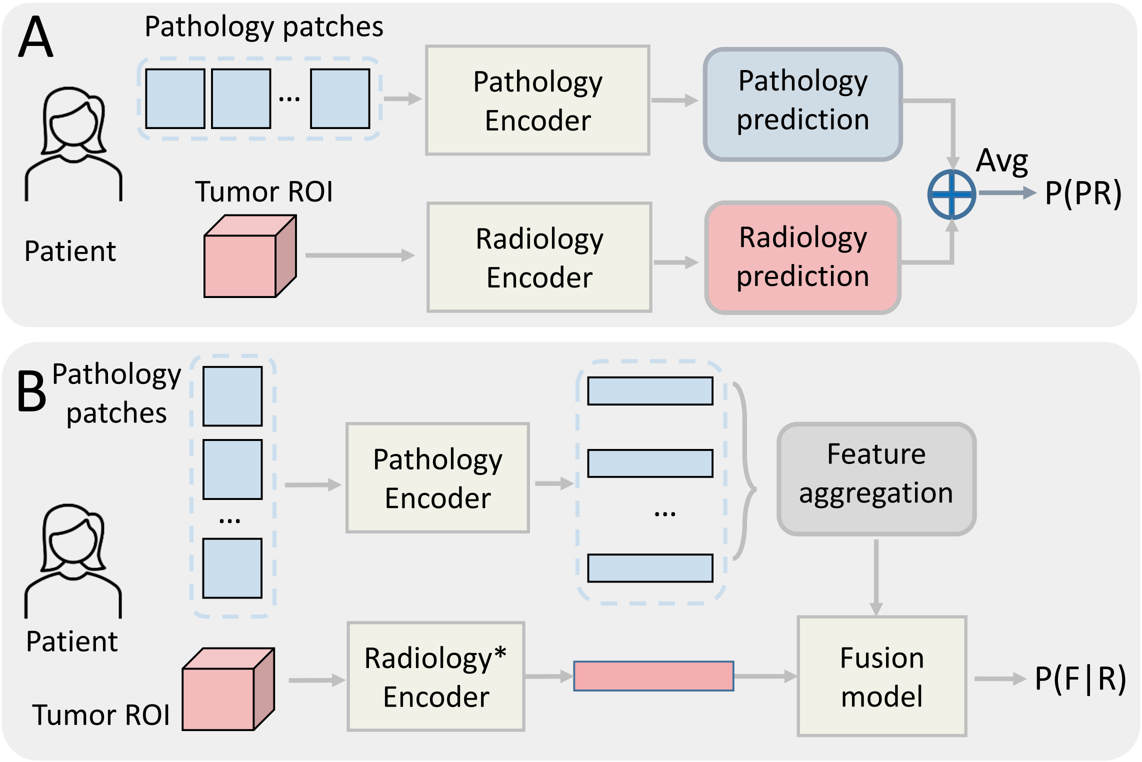


Figure 3.13: Baseline experiments on multimodal fusion. A. Decision level multimodal fusion, $\mathscr{P}(P_{ath}R_{ad})$ in Eq.3.5. B. Radiology-guided feature-level fusion, probability distribution follows $\mathscr{P}(F_{ea}|R_{ad})$. '*' means the model is well-trained and frozen in pipeline training.

3.4.2.3 Model prediction fusion on multiple levels

3.4.2.3.1 Decision level multimodal fusion

Fig. 3.13-A shows the decision level fusion. Both models are trained and make the prediction separately. The mean of predicted probability from pathology and radiology is taken as the MSI prediction score for the patient. Based on the well-trained uni-model on pathology images and radiology data, the decision-level multimodal fusion employs the patient-level MSI prediction for the final decision. From the well-trained pathology uni-model, the pathology image W^i from patient i has predicted MSI probability P_p^i. Similar to pathology prediction, radiology CT scans C^i from patient i can get MSi probability prediction P_r^i. The decision level fused prediction follows $P^i = (P_r^i + P_p^i)/2$.

3.4.2.3.2 Feature level multi-modal fusion

Fig. 3.13-B shows the model fusion on the feature level. The feature embedding abstracted from pathology patches is aggregated as a single feature representing the bag of cropped pathology patches. Each pathology patch is generated as an embedding e^i from patch x^i. The generated embedding $e^i \in \mathbb{R}^{1\times 512}$ is not representative of the WSI. We first aggregate e^i when $i \in [1,N]$ to a single feature for further feature-level fusion. Referring to the Multi-instance Learning (MIL) methods(Raju et al., 2020), we use maxing pooling on each channel of embeddings to aggregate the single patches embedding to patient pathology embedding e. The aggregation process follows Eq.3.9 where $d \in [0,511]$ and $e \in \mathbb{R}^{1\times 512}$.

$$e_d = max_{i=0,\ldots,N} e_d^i \tag{3.9}$$

Radiology feature embedding is abstracted from segmented tumor ROI. The feature embeddings from both modalities are fused by feeding into the fusion model. Two major feature-level fusion strategies are investigated in our study, the Transformer-based or MLP-based fusion model. Transformer model (Dosovitskiy et al., 2020) takes the aggregated WSI feature embedding and radiology ROI embedding as input. Following the standard approach in the transformer model, a learnable class token is added to the input embedding sequence. Multi-layer Perceptron (MLP) fusion model concatenates embeddings from two modalities and is then finetuned with the patient MSI label. The dim of two modality embeddings are both 1×512.

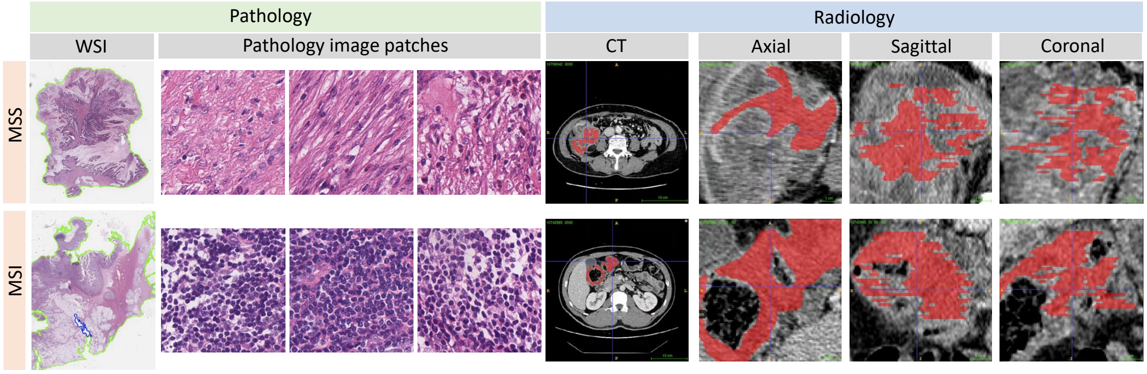


Figure 3.14: Data visualization of the dataset. First row shows two modalities image from MSS subjects. The second row shows data from MSI subject.

3.4.3 Experiments

3.4.3.1 Dataset

We collect an in-house dataset that has the paired pathology WSIs and CT images from 352 patients shown in Fig.3.14. The dataset includes 46 MSI patients and 306 MSS patients. The venous phase is used for tumor annotations by a board-certified radiologist with 14 years of specialized experiences. The median imaging spacing is $0.76 \times 0.76 \times 5$ mm^3. The pathology WSI is at a gigapixel level maintained in a pyramid structure. Each level each layer contains a reduced-resolution version of the image from 5×, 10×, and 40× magnification. The highest level of the pyramid is the full-resolution image which is 40× in 0.25 μm per pixel. The image patches are 448×448 cropped from 40× level and resize to 224×224.

To thoroughly evaluate the dataset performance, we use 5-fold cross-validation in all model evaluations. Since the MSI/MSS ratio is unbalanced, the MSI patients and MSS patients are evenly split into five folds to guarantee a fair MSI/MSS ratio in each fold. For each experiment, three folds of data are used for training, one fold for validation, and the rest one fold for testing. By picking up different folds as testing data, five-set experiments are conducted. The average AUC score is used as the evaluation criterion.

3.4.3.2 Experimental Design

In the experiments, we aim at evaluating the proposed Bayesian-based multimodal multi-level fusion model. The experiment parts verify two research questions: (1) whether multimodal fusion provides better performance over the uni-model (rely on single data modality), (2) if our proposed Bayesian-based model $\mathscr{P}(F_{ea}P_{ath}R_{ad})$ achieves the optimal fusion strategy over other fusion models. The ablation study is explored feature aggregation and feature-level fusion strategy.

Pathology uni-modal prediction The uni-model on pathology data is separated into two steps. First, the WSIs are cropped by the CLAM model into 224×224 patches. The patches use the WSI labels in model training. ImageNet-pretrained ResNet-18 is trained for 100 epochs and the batch size is set to 128. In the testing stage, the average probability of patches from the same WSI is used as patient WSI probability prediction. The final model performance is the average score of 5 testing fold.

Radiology uni-modal prediction For the Radiology uni-model, we construct the training data by selecting six essential slides based on CT image and annotated tumor region. Only one ROI block is cropped from each CT and constructs the six-channel training data (batch size = 2). ImageNet pre-trained ResNet-18 is employed as the encoder.

Decision level fusion prediction Different from uni-model training from scratch, decision-level fusion is based on a well-trained uni-model. Based on the 5-fold well-trained model, we feed the test fold data to the corresponding trained model and get the MSI prediction by pathology data. The same process goes for radiology data. The decision-level fused prediction is computed by average MSI probability from two modalities.

Feature-level fusion prediction

Instead of fusing the probability prediction from two modalities, the regular feature level fusion model fuses the embeddings generated from the two modalities' encoders. Both modality encoders are trained from scratch. For the radiology-guided feature level fusion, two modalities of data and a well-trained radiology uni-model are needed. The pathology data is fed into an end-to-end training path. The output of the pathology path is an aggregated feature for pathology WSI. The radiology path is an abstracted feature by pre-trained radiology uni-model from its corresponding training model. For a patient sample, two 1×512 features from pathology and radiology data are fed into fusion model. For the Transformer-based model, we choose ViT-S as our backbone. Our ViT-S model depth is 8, the head number is 12. Multi-layer perception (MLP) hidden feature dimension is 1024. The input matrix is in 3×512. CNN-based feature level fusion concatenates the feature from two modalities into one feature with a length of 1024. An MLP is constructed to map the concatenated feature to the final fusion prediction, which has two fully connected layers when the hidden dimension is 256.

Bayesian-guided multi-level fusion prediction

For the Bayesian-guided fusion model, we used the same input data as previous fusion experiments: a bag of pathology image patches and radiology CT tumor Region of Interest (ROI). The patient MSI prediction from radiology can be generated by the pre-trained model. The feature abstracted from radiology ROI can be generated from the second last layer's output. The feature and patient-level prediction from pathology follow the same procedure as radiology except the pathology encoder is trainable. The fusion model we used is ViT-S for the Transformer-based model and a two-layer MLP for MLP based fusion model. The average score of the pathology, radiology, and feature fusion MSI probability prediction is used as the final prediction.

3.4.4 Results

We conduct experiments on 5-fold cross-validation and model performances are shown in Table. 3.7. Our proposed multi-level multi-modality fusion pipeline is compared with the single-modality model and fusion methods. From the average AUC score across 5-fold experiments, the performance of unimodal relies on pathology image and radiology image are 0.6847 and 0.7348, respectively. The decision-level fusion has an average AUC score of 0.7908 which outperforms unimodal prediction score. The feature-level fusion model shows better performance by using Vision Transformer than MLP. Without radiology guidance, feature-level fusion model (avg AUC: 0.7289) performs better than pathology unimodal but worse than radiology unimodal. The radiology data can guide feature-level fusion model training by getting AUC score of 0.7696 better than 0.7289. Radiology-guided feature-level fusion model shows better performance than feature-level fusion without a guide. By combining the decision-level and feature-level information from two image modalities, our proposed multi-level multi-modality pipeline get the best AUC 0.8177 over the rest of MSI CRC strategies.

Table 3.7: AUC on MSI prediction

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Patho unimodal(Yamashita et al., 2021)	0.6502	0.7282	0.8530	0.8819	0.6500	0.6847
2.5D Radio unimodal	0.5615	0.8333	0.7520	0.7163	0.8158	0.7348
Decision-level fusion	0.6956	0.8313	0.8536	0.8948	0.6785	0.7908
Feature-level fusion	0.619	0.6528	0.7698	0.7083	0.6730	0.7289
Radio-guided feature fusion	0.7218	0.7558	0.7698	0.7678	0.8127	0.7696
M^2**Fusion**	0.8278	0.8055	0.7341	0.8989	0.8222	**0.8177**

An ablation study on exploring the pathology feature aggregation strategy and multimodal feature level fusion backbone is shown in Table.3.8. The combination of average pooling on pathology feature aggregation and using a Transformer as feature-level fusion backbone has the best AUC performance.

3.4.5 Conclusion

We proposed a multi-level multi-modality fusion pipeline for colorectal cancer MSI status prediction based on pathology WSIs and CT images. We introduce Bayes' theorem to fuse the information

Table 3.8: Ablation study for pathology feature aggregation and feature-level fusion strategy

Feature aggregation	Feature fusion	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Conv	Transformer	0.5423	0.6012	0.7976	0.7540	0.7746	0.6939
Avg	CNN	0.5786	0.7004	0.7202	0.7044	0.7333	0.6874
Conv	CNN	0.6593	0.7321	0.7599	0.6706	0.7047	0.7053
Avg	Transformer	0.7218	0.7758	0.7698	0.7678	0.8127	**0.7696**

from two image modalities on both the feature level and decision level. The experiment result shows (1) radiology and pathology image fusion (decision level fusion) helps CRC MSI prediction by combining the two modalities' information from the same patient, and (2) radiology-guided feature-level training outperforms the model that directly fuses two modalities' features. Our Bayesian-based fusion on both decision-level and feature-level achieves the best performance.

3.5 Multi-Level Text-Guided Representation End-to-End Learning for Whole Slide Image Analysis

3.5.1 Introduction

Analyzing Whole Slide Images (WSIs) is a critical aspect of medical imaging research. WSIs are digitized scans of histological samples captured at multiple magnifications, preserving both the overarching view and intricate microscopic details. From a broader perspective, WSIs offer a macroscopic overview of tumor distribution throughout the entire digital slide. This allows for the study of spatial relationships and general tumor traits. At the same time, WSIs enable detailed inspections of cell and tissue structures at the microscopic scale, significantly enhancing diagnostic accuracy and advancing the field of histopathology research (Pantanowitz et al., 2011).

Over recent years, computer vision has become increasingly crucial and successful in analyzing Whole Slide Images (WSIs), tackling the challenges of handling their super high-resolution ($> 10^9$ gigapixels) for tasks like image classification, object detection, and segmentation. Unlike these tasks, our study confronts the unique challenge of performing multi-modal representation learning (involving both image and textual data) for WSIs, focusing on modeling information across multiple scales in both images and text. Textual data in this context can describe both broad and detailed aspects of multi-scale WSIs. A key unresolved question is how to effectively learn representations that encompass both global and local features.

Current methods often rely on manually labeling areas of interest for local representation or using multi-stage learning to merge these local features into a global representation. Yet, these approaches typically fall short of seamlessly integrating multi-scale image representations with text data in an end-to-end process.

In this paper, we propose Multi-Level Text-Guided Representation End-to-End Learning (mTREE), an innovative text-guided method that effectively captures multi-scale image representations through the use of accompanying textual pathology data. mTREE uniquely blends two formerly separate processes – the identification of crucial areas (**"global-to-local"**) and the creation of a WSI-level image-text representation (**"local-to-global"**) – into a unified, end-to-end learning framework.

While text-based clinical records are consistently available, they haven't been fully utilized in multi-modal representation learning for high-resolution images. Our goal is to develop an algorithm that leverages these textual records to guide the selection of diagnostic patches and aggregate WSI

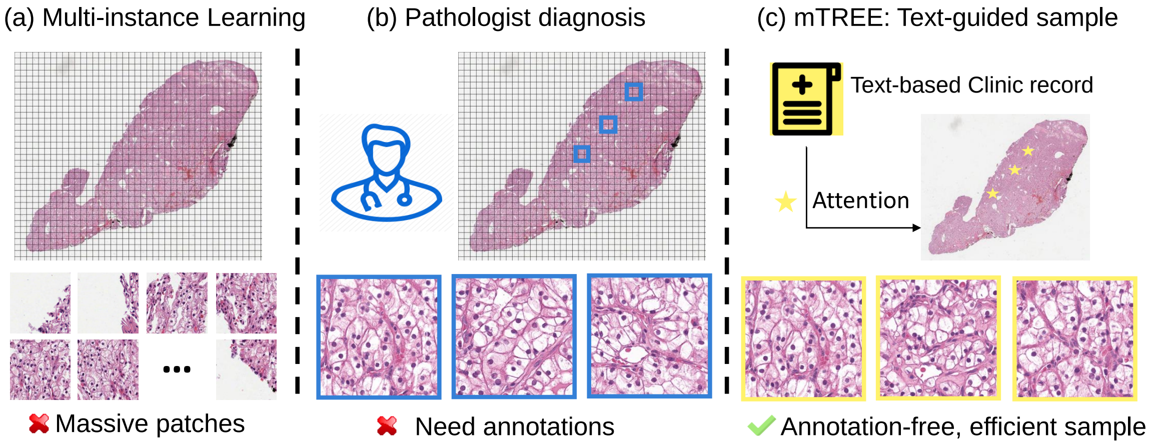

(a) Multi-instance Learning
(b) Pathologist diagnosis
(c) mTREE: Text-guided sample
Text-based Clinic record
Attention
✗ Massive patches
✗ Need annotations
✓ Annotation-free, efficient sample

Figure 3.15: Comparison between multi-instance learning, pathologist diagnosis, and our proposed mTREE. (a) Traditional multi-instance learning needs to process all patches without patch selection. (b) Pathologists in the diagnosis process focus on the most essential patches selected by manual efforts. (c) Our proposed mTREE generates text-guided attention to sample efficiently without manual annotation.

representations without the need for manual annotations. We hypothesize that there is an inherent correlation between the text and image domains; for a given WSI, the clinical text can provide criteria for selecting WSI patches and extracting features.

Mathematically, Given a WSI X, considered as a set of image patches $\{x_i\}_{i=1}^N$ where $x_i \in X$, and a corresponding label Y for WSI X, our approach involves training two mappings. The first is for patch selection $A : \{x_i\}_{i=1}^N \to \{x_i\}_{i=1}^K, (K \ll N)$, and the second is for feature abstraction $E : \{x_i\}_{i=1}^K \to Y$. Textual data provides guidance for both mappings: optimizing the selection from the original WSI set and consolidating patch-level features into a comprehensive WSI feature. This dual mapping requires a multi-level approach for text-guided analysis, executed in an end-to-end fashion. In the first mapping, patch selection from $\{x_i\}_{i=1}^N$ depends on each patch's relevance to the final prediction, determined by an attention map that scores each patch's importance. Due to the extensive size of WSIs, the attention map is initially learned on lower-resolution images and then mapped to high-resolution images based on coordinate relationships. This approach allows the model to process only a fraction of patches (K out of N) when $K \ll N$. In the second mapping, the model uses the features extracted from the selected patches, with the text feature identifying and amalgamating the most pertinent features (those with smaller feature distances) into a unified WSI representation.

In summary, our study introduces a text-guided representation learning method aimed at improving efficiency and extracting features from vital image regions, thereby eliminating the confusion caused by unnecessary image patches. Our approach does not require image annotations from pathologists. Instead, it leverages text descriptions from clinical records to guide the learning process of WSI representations at multiple levels in an integrated, end-to-end manner. We have applied our method to various applications, including image classification and survival prediction across multiple WSI datasets, and have compared it with previous approaches based on Multiple Instance Learning (MIL) models.

The key contributions of our work are fourfold:

- We present the first efficient visual-language model for gigapixel WSIs, operating in a seamless end-to-end fashion.
- We utilize text information to optimize learning strategies across multiple levels.
- Our pipeline is weakly supervised at the WSI level, eliminating the need for patch-level annotations from pathologists.
- Our model offers explainability by providing visualizations at different levels, such as attention maps and significant patches.

3.5.2 Related Work

3.5.2.1 Multi-instance learning

For pathology image analysis, Multi-Instance Learning (MIL) has emerged as a prominent paradigm, offering a robust framework to address challenges associated with the lack of patch-wise annotation of pathological images (Yao et al., 2020; Hou et al., 2016b; Lu et al., 2021a). Different from the supervised learning method on patches, (Edwards and Storkey, 2016; Zaheer et al., 2017) regard the pathology image as a collection of multiple instances or regions, each potentially harboring critical information for diagnostic or prognostic purposes. This approach allows the model to operate on bags of instances for weakly-supervised learning (Carbonneau et al., 2018). MIL has demonstrated its efficacy in capturing nuanced spatial relationships (Zhao et al., 2020b) and patterns (Wu et al., 2022) within pathology images, accommodating the inherently diverse nature of tissue structures (Shao et al., 2023; Campanella et al., 2019) and cellular compositions (Kraus et al., 2016).

Upon the set-based concept, Ilse et al. (Ilse et al., 2018) apply the attention mechanism to Whole Slide Images (WSIs). In a similar vein, Yao et al. (Yao et al., 2020) integrated attention-based MIL into clustered phenotypes, yielding promising outcomes. Furthermore, (Campanella et al., 2018) validated the MIL performs well on large-scale WSI datasets.

3.5.2.2 Attention sampling

Performing analysis on large images, attention sampling (Xu et al., 2015a) has emerged as a powerful technique to efficiently process extensive visual datasets by selectively focusing computational resources on regions of interest. Attention sampling aims to address the challenges posed by the sheer scale of high-resolution images, where the majority of the content may be irrelevant to the specific task at hand (Zheng et al., 2019). Attention sampling leverages mechanisms inspired by human visual attention (Hassanin et al., 2022), directing computational resources toward salient regions while bypassing less informative areas. Notable approaches include the integration of attention mechanisms within convolutional neural networks (CNNs) (Wang et al., 2023; Xue et al., 2022; Wang et al., 2021a) to dynamically weigh the importance of different image regions. Additionally, attention sampling strategies, such as region-based methods and attention-guided sampling, have been proposed to enhance computational efficiency in tasks such as object detection (Cheng et al., 2021; Li et al., 2019), image classification (Dong et al., 2022; Wang et al., 2017), and segmentation (Kulharia et al., 2020; Shi et al., 2022).

3.5.2.3 Visual language model in WSI analysis

The integration of visual language models (VLM) has emerged as a cutting-edge approach, revolutionizing the interpretation of large-scale pathological images. Visual language models combine the strengths of natural language processing (NLP) and computer vision, enabling a comprehensive understanding of complex visual information assisted by knowledge from multiple domains. Unlike fine-tuning, VLM is based on prompt prediction in the template, as seen in CLIP (Radford et al., 2021) and CoOp (Zhou et al., 2022). The trained language model has a strong capability in knowledge and zero-shot learning (Brown et al., 2020). By leveraging pre-trained language models such as BERT (Devlin et al., 2018) and adapting them to the unique challenges of WSI, researchers have achieved remarkable strides in capturing contextual relationships and hierarchical structures within pathology images (Huang et al., 2023). These models empower the extraction of meaningful features and semantic understanding, enhancing the interpretability of WSIs for tasks such as image classification, tumor detection, and prognosis prediction (Lu et al., 2023).

3.5.3 Methods

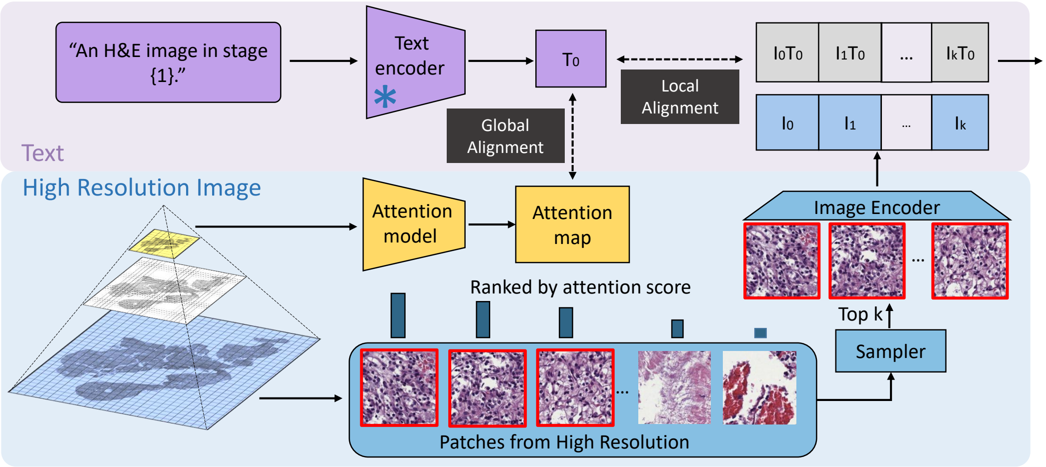


Figure 3.16: This figure demonstrates the proposed mTREE pipeline. The upper panel shows the text process flow. The text encoder is frozen with pre-trained weights. Text feature T_0 is used for global alignment and alignment of image patch features. The lower panel shows the WSI analytic flow. The attention model learns an attention map from the WSI in low resolution. The attention map aligns with the text feature T_0. The image patches tiled up from high-resolution WSI are ranked by attention score. The image features $I_0, I_1 \ldots I_k$ abstracted from image patches with higher attention scores are aggregated with text feature T_0.

As illustrated in 4.11, our goal is to learn the multi-level mapping from gigapixel WSI X to the representation R guided by text prompt T. To better describe the model learning on gigapixel images, we regard high-resolution images (base layer in image pyramid in 4.11) as a collection of image patches $\{x_i\}_{i=1}^{N}$ $(x_i \in X)$. To efficiently learn the WSI representation, the global-level alignment learns an attention map M from WSI in low resolution W_{low}. The global alignment on M is between image attention map M_w and text attention map M_t. The image patch collection $\{x_i\}_{i=1}^{N}$ is ranked by the attention score in attention map M. Top K image patches $\{x_i\}_{i=1}^{K}$ with highest attention score are selected by the sampler. Image patch features $\{I_i\}_{i=1}^{K}$ is abstracted by image encoder E from top K image patches $E(x_i)$. The image feature I_i is selected by the cosine similarity with T_0. The WSI representation R is aggregated by the I_i weighted by distance $I_0 \times T_0$. At the core of our approach is the idea of multi-level alignment from supervision contained in natural language

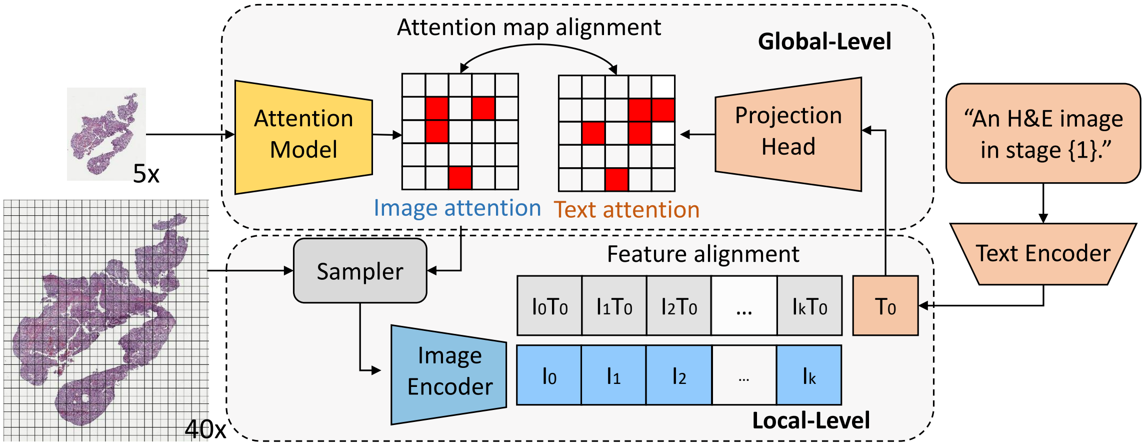


Figure 3.17: This figure presents the principle of multi-level text guidance. Global-level text guidance (upper panel) aligns the attention map from images and text. Image attention map is learned from low-resolution WSI, while text attention is projected from the text feature T_0. Local-level text guidance (lower panel) performs patch selection by computing the cosine similarity distance to the text feature T_0 and aggregates features from both image and text.

and the end-to-end training manner to coordinate the region localization and feature extraction.

3.5.3.1 Global alignment

The attention map can, in theory, learn the significance of the region playing a role in the final prediction. However, for the high-resolution WSI normally in size of $100,000 \times 100,000$, a gigapixel attention map is too large for attention model capacity to learning. Considering the histopathology image composed of regions of different tissue types, the region distribution of WSI makes attention map learning even harder to match the target prediction. Since learning the attention map for the high-resolution image is not valid, we can learn the attention map in low-resolution WSI to reduce the training burden and localize the rough essential region. To further optimize the attention map, the global-level alignment is between text feature T_0 and low-resolution WSI X_{low}. The attention mappings on the text side $A_t : T_0 \rightarrow M_t$ and WSI side $A_w : X_{low} \rightarrow M_w$ provides the WSI attention map M_w and the text attention map M_t. The mapping A_t is achieved by a projection head, composed of multiple fully connected layers. We use MSELoss as attention loss $\mathscr{L}_{Attn}$ to match the M_w with M_t as expressed in 3.10.

$$\mathscr{L}_{Attn} = \frac{1}{N}\sum_{i=1}^{N}(M_w^i - M_t^i)^2 \tag{3.10}$$

To prevent the attention from collapsing and to encourage non-zero values in the attention map, the sparsity-inducing regulation term $\mathscr{L}_{Sparse}$ is applied to both M_t and M_w, penalizing overly sparse attention maps following 3.11.

$$\mathscr{L}_{Sparse} = \frac{1}{N}\sum_{i=1}^{N}|M_t| + \frac{1}{N}\sum_{i=1}^{N}|M_w| \tag{3.11}$$

To keep the low-resolution WSI in the same size from the same magnification, the original image is center-cropped before being fed into the attention model. To prevent the attention from focusing only on the cropped image boundary, the $\mathscr{L}_{Boundary}$ is defined by gradients of the attention map with respect to the x-axis (G_x) and y-axis (G_y). The $\mathscr{L}_{Boundary}$ is expressed in 3.12 where $G_x(M_w)_i$ is the i-th element in the gradient of M_w with respect to x-axis and $G_y(M_w)_i$ is the i-th element in the gradient of M_w with respect to y-axis.

$$\mathscr{L}_{Boundary} = \lambda\left(\sum_{i=1}^{N}|G_x(M_w)_i| + \sum_{i=1}^{N}|G_y(M_w)_i|\right) \tag{3.12}$$

The full loss function for attention map global alignment is:

$$\mathscr{L}_{Global} = \mathscr{L}_{Attn} + \mathscr{L}_{Sparse} + \mathscr{L}_{Boundary} \tag{3.13}$$

3.5.3.2 Local alignment

The attention map M provides the criterion score for image patch significance. The top K patch selected from $\{x_i\}_{i=1}^{N}$ provides the most essential patches for the prediction. As shown in 3.17 lower panel, essential patch collection $\{x_i\}_{i=1}^{K}$ selected by the sampler is fed into the image encoder to generate the image feature $\{I_i\}_{i=1}^{K}$. Because the image patches and text are both encoded into feature space, the similarity of text feature T_0 with each image feature candidate I_i can be computed and ranked based on 3.14

$$\text{Cosine Similarity}(\{I_i\}_{i=1}^{K}, T_0) = \left\{\frac{I_i \cdot T_0}{\|I_i\|\|T_0\|}\right\}_{i=1}^{K} \tag{3.14}$$

Image feature $\{I_i\}_{i=1}^{J}$ are selected based on the cosine similarity with the text feature T_0. The final representation for WSI is aggregated with the selected image features, as shown in 3.15.

$$R = \sum_{i=1}^{J} \left(\frac{I_i \cdot T_0}{\|I_i\| \|T_0\|} \cdot I_i \right) \tag{3.15}$$

3.5.3.3 End-to-end training with image sampler

The multi-level alignment under text guidance is built in an end-to-end manner to provide the advantages from two aspects: (1) Coordinate the optimization on the attention map and feature selection and aggregation. (2) Provide the backpropagation path from the final prediction to the attention map for the sampler. Representation learning with attention maps often involves a two-stage training process, where the first stage focuses on learning a representation, and the second stage incorporates attention mechanisms as in (Xu et al., 2014; Vaswani et al., 2017; Mnih et al., 2014; Ng et al., 2015). Insufficient end-to-end optimization potentially leads to suboptimal integration of attention and representation learning. To better describe the end-to-end manner, the pseudocode of mTREE is provided.

Algorithm 1 Pseudocode for mTREE implementation

```
Input: Gigapixel WSI $X$, Text Prompt $T$
Output: WSI Representation $R$
Learn attention map $M_w$ from low-resolution $X_{\text{low}}$
Align image attention map $M_w$ and text attention map $M_t$ with attention loss
Rank image patches $\{x_i\}_{i=1}^{N}$ using attention score in $M_w$
Select top $K$ image patches $\{x_i\}_{i=1}^{K}$ with highest attention score
Extract image features $\{I_i\}_{i=1}^{K}$ with image encoder $E$
Select image feature $I_i$ based on cosine similarity with $T_0$
Aggregate WSI representation $R$ using $I_i$ weighted by distance $I_0 \times T_0$
```

3.5.4 Experiments

3.5.4.1 Data description

To substantiate our proposed text-guided representation learning approach, integrating histological and text features, we sourced glioma and clear cell renal cell carcinoma data from the TCGA, a comprehensive cancer data consortium housing paired high-throughput text in clinic records and diagnostic whole slide images. This dataset is enriched with ground-truth survival outcomes and histologic grade labels. For both the TCGA-KIRC (519 WSIs) and TCGA-GBMLGG (1589 WSIs) projects, region-of-interests (ROIs) from diagnostic slides are provided by (Chen et al., 2020a). For clear cell renal cell carcinoma in the TCGA-KIRC project, 512×512 ROIs from diagnostic whole

Method	Inputs	Patch #	Image encoder	Acc	C-Index
PathomicFusion	diagnostic regions (**manual**)	20	ResNet-50	N/A	63.1
AttenDeepMIL	diagnostic regions (**manual**)	20	ResNet-50	60.9	61.5
AttenSample	raw WSI (**automatic**)	1	ConvNet	49.1	55.4
AttenDeepMIL	raw WSI (**automatic**)	100	ResNet-50	51.0	58.8
CLAM	raw WSI (**automatic**)	>5000	ResNet-50	57.5	60.1
mTREE (Ours)	raw WSI (**automatic**)	10	ResNet-50	63.1	63.2
mTREE (Ours)	raw WSI (**automatic**)	20	ResNet-50	**64.7**	**65.1**

Table 3.9: Cancer grade classification and survival prediction results on KIRC dataset. Acc represents the accuracy of grade classification, while the C-Index evaluates the survival prediction performance.

slide images are provided as the diagnostic region. This yielded 3 ROIs per patient (512×512 at 40× magnification) for 417 patients, resulting in a total of 1251 images. For the TCGA-GBMLGG project, 1024×1024 region-of-interests (ROIs) from diagnostic slides are leveraged. The WSI data is publicly available on the TCGA database (Tomczak et al., 2015a).

3.5.4.2 Data preprocessing

Both WSI image and text data require preprocessing before feature extraction. The preprocessing for both datasets follows the same strategy.

WSI image data. The input image data for our pipeline is provided from two levels: low-resolution images and high-resolution images. The low-resolution images are from 5x magnifications in the WSI pyramid structure. To ensure low-resolution images in the same size and scale, we center-crop the $5,000 \times 5,000$ patches from the low-resolution images. All $5,000 \times 5,000$ patches are then resized to 500×500.

Text data. Follow the design in (Huang et al., 2023; Lu et al., 2023), text information is composed of templates and prompts curated from the clinical records. In our experiments, the paragraph related to ”survival time” and the ”cancer grade” are used as text information.

3.5.4.3 Network architectures

We adopt the representation learning flow from the ”low-resolution” to ”high-resolution”, as proposed in Attention-sampling (Katharopoulos and Fleuret, 2019), which has shown impressive results on megapixel image analysis. Based on the ”low-resolution” to ”high-resolution” strategy, our mTREE pipeline is composed of three parts: image feature analysis models, text feature analysis models, and alignment blocks between image and text.

Image feature analysis model. Attention model for low-resolution WSI is a Convolution Network (ConvNet) with four convolution layers. We use 3×3 convolution kernel and the channel number of four convolution layers is [8, 16, 32, 1]. The Sampler ranks image patches and selects the top K patches. No learnable parameters in the sampler. ResNet-50 with ImageNet pre-trained weight is used as the image encoder.

Text feature analysis model. The text encoder is the pre-trained ViT-B/32 used in CLIP (Radford et al., 2021). The text encoder is frozen in the training process.

Alignment block. The global alignment block includes a projection head composed of two sequential convolution layers to project text feature to text attention map. The local alignment between text features and image features is based on the cosine similarity matrix in the shape of $1 \times K$.

Method	Inputs	Patch #	Image encoder	Acc	C-Index
PathomicFusion	diagnostic regions (**manual**)	20	ResNet-50	N/A	**72.4**
AttenDeepMIL	diagnostic regions (**manual**)	20	ResNet-50	78.8	71.4
AttenSample	raw WSI (**automatic**)	1	ConvNet	70.4	65.4
AttenDeepMIL	raw WSI (**automatic**)	100	ResNet-50	70.6	63.6
CLAM	raw WSI (**automatic**)	>5000	ResNet-50	75.3	65.7
mTREE (Ours)	raw WSI (**automatic**)	10	ResNet-50	76.5	69.0
mTREE (Ours)	raw WSI (**automatic**)	20	ResNet-50	**79.6**	70.1

Table 3.10: Cancer grade classification and survival prediction results on GBMLGG dataset. 'Acc' represents the accuracy of grade classification, while the C-Index evaluates the survival prediction performance.

3.5.4.4 Training details

We apply our mTREE to two TCGA datasets (KIRC and GBMLGG) on two downstream tasks: grade classification and survival prediction.

Tasks. The KIRC dataset has three grades (Stage I, Stage II, and Stage III) for grade classification. The patient's overall survival time in month is used as the label for the survival prediction task. The GBMLGG dataset also has three grades (2, 3, 4) for grade classification. The ”Time to last follow-up or death (month)” is used as the label for survival prediction in the GBMLGG dataset.

Hyper-parameters. The three most important parameters are evaluated for our proposed mTREE. The first one is the size of the attention map learned in global alignment. Tuned by the projection head channel number and attention model structure, we evaluated attention map size in 123×123 and 246×246. The second parameter is the sample number from the attention map (K in 3.14). Based on the number of patches from the diagnostic region provided by (Chen et al., 2020a), approximately 20 patches for each WSI, we evaluated sample number in set 5, 10 20, 50. The third parameter is the sample number from the image patch features (J in 3.15). According to $K \in \{5,10,20,50\}$, we sampled $J \in \{2,5,10,20\}$.

Metrics. The metric used for grade classification tasks is accuracy (ACC). The ACC evaluates the WSI representation performance on class prediction tasks with discrete labels.

The metric for the survival prediction task is the C-Index. It quantifies the concordance between predicted and observed survival times, with a higher C-index indicating improved predictive accuracy.

3.5.4.5 Baseline experiments for comparison

To validate the advantages of the proposed mTREE pipeline, the baseline experiments comparisons are compared from three aspects: (1) MIL-based model: we use AttenDeepMIL (Ilse et al., 2018) as a general MIL-based model and CLAM (Lu et al., 2021b), designed specifically for WSI analysis. (2) Attention Sampling (Katharopoulos and Fleuret, 2019), and (3) PathomicFusion (Chen et al., 2020a).

AttenDeepMIL. The implementation of AttenDeepMIL follows the settings in (Ilse et al., 2018). ResNet-50 with ImageNet-pretrained weights is used as the image encoder. For both the KIRC dataset and the GBMLGG dataset, two patch selection strategies are evaluated: (1) diagnostic region (DR), and (2) tiled-up image patches from 40x WSI (Origin).

CLAM. In (Lu et al., 2021b), CLAM provides the implementation of MIL for classification. For a fair comparison, the image encoder is ResNet-50, similar to other baseline methods. CLAM processes all image patches from WSI, except the background patches, normally more than 5,000 patches for a WSI.

AttenSample. Attention sampling input set has an image in high resolution $(1,500 \times 1,500)$ and a low-resolution image rescaled by a ratio of 0.1 (150×150). For the TCGA dataset, the high-resolution image is from a 5x magnification WSI and center-cropped in size of $5,000 \times 5,000$. The low-resolution image is rescaled by a ratio of 0.1 to 500×500.

PathomicFusion. In (Chen et al., 2020a), PathomicFusion is a multi-modal fusion method incorporated with image, genomics, and cell graph data. In our experiments, only image data is used for performance comparison.

3.5.5 Results

Following the experiment settings, two datasets TCGA-KIRC and TCGA-GBMLGG on grade classification and survival prediction are discussed. To optimize the hyperparameter settings, an ablation study of different parameters is evaluated on the survival prediction task on the GBMLGG dataset. To provide the model with explainability, the visualization of the attention map and selected diagnostic region by mTREE is presented.

3.5.5.1 KIRC

In 3.9, we compare the performance of the proposed mTREE with baselines on the TCGA-KIRC dataset for the grade classification task. It is observed that the performance of MIL-based methods improves with an increasing sample number from the WSI patch collection. When processing all patches from the WSI, CLAM achieves an accuracy of 57.5%. Remarkably, our proposed mTREE outperforms, achieving a superior accuracy of 64.7% with just 20 sampled patches from the WSI, surpassing even AttenDeepMIL with a diagnostic region.

Moving on to the survival prediction task in 3.9, we observe a performance trend similar to the grade classification task. MIL-based methods exhibit better performance with a diagnostic region compared to original image patches. The best performance from the MIL baseline achieves a C-Index of 0.631 when trained on the diagnostic region. Notably, our proposed mTREE demonstrates superior performance (C-Index 0.651) over the baselines, utilizing both original image patches and diagnostic regions.

3.5.5.2 GBMLGG

In Table.3.10, we present the prediction performance of the proposed mTREE and baselines on the grade classification task. Similar to the performance comparison in the grade classification task,

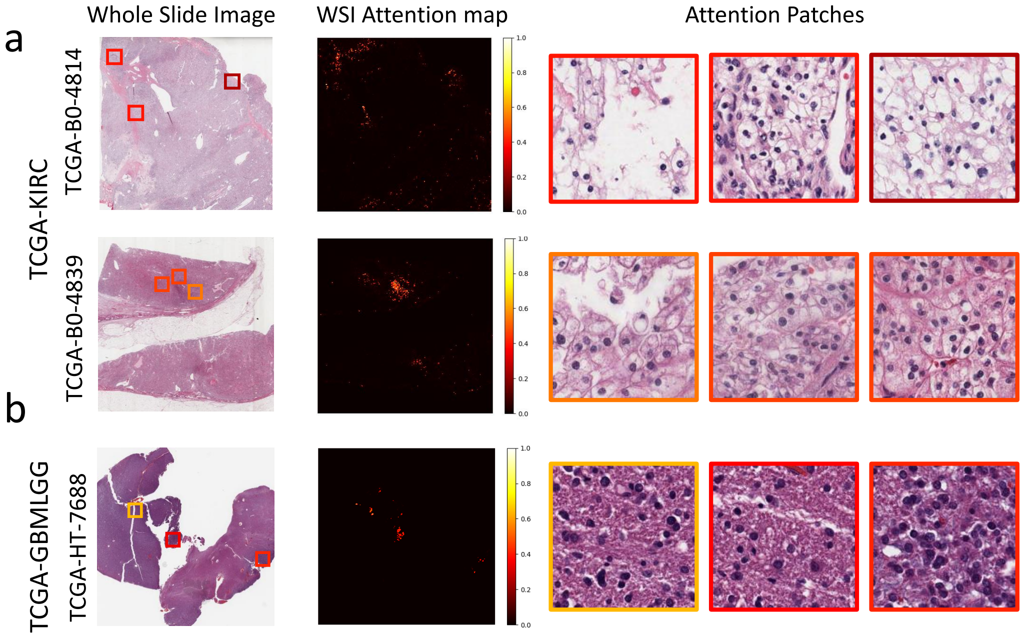


Figure 3.18: This figure presents the visualization of WSI-level attention and the automatically derived diagnosis patches. For WSIs in the TCGA-KIRC dataset and TCGA-GBMLGG dataset, the attention map (middle panels) is learned from WSI (left panels), highlighting essential tissue regions. Essential image patches (right panels) are selected according to the attention score. The image boundary color indicates the according attention score.

MIL-based methods exhibit better performance with a diagnostic region compared to original image patches. The best performance from the MIL baseline achieves an accuracy of 78.8% when trained on the diagnostic region. Notably, our proposed mTREE outperforms the baselines, achieving a superior accuracy of 79.6% with both original image patches and the diagnostic region.

In Table.3.10, we present the prediction performance of the proposed mTREE and baselines for the survival prediction task. The MIL-based method exhibits better performance with a diagnostic region than with original image patches. The best performance from the MIL baseline achieves a C-Index of 0.724 when trained with the diagnostic region. Notably, our proposed mTREE outperforms the baselines, achieving a better performance (C-Index 0.701) with original image patches.

3.5.5.3 Evaluation for multi-level text alignments

In this section, we compare the performance of the proposed mTREE with and without global and local alignment. The results are presented in Table.3.11. From the performance shown in Table.3.11, both global and local alignment contribute to performance improvement in all four tasks. However,

Method	Global align	Local align	KIRC ACC	KIRC C-Index	GBMLGG ACC	GBMLGG C-Index
			49.1	55.4	70.4	65.4
mTREE	✓		51.0	56.3	70.7	65.2
	✓	✓	**64.7**	**65.1**	**79.6**	**70.1**

Table 3.11: Ablation study for multi-level text alignments is shown in this table. The accuracy of grade classification (ACC) and the survival prediction performance (C-Index) are presented.

global alignment provides a limited contribution (as shown in the second row of Table.3.11). From another perspective, the coordination between local and global alignment underscores the advantages of an end-to-end training approach.

3.5.5.4 Visualization

The attention map (Figure.3.18 middle) obtained through global alignment serves as a crucial tool for improving the interpretability of mTREE in the context of whole-slide image (WSI) analysis. The heightened intensity in the attention map accentuates key regions within the WSI that significantly contribute to the final prediction. In the domain of weakly-supervised learning, these bright regions indicate areas of essential diagnostic relevance. To enhance human understanding, image patches identified as having high diagnostic importance are presented in a zoomed-in view (Figure.3.18 right). The color of the image boundary indicates the corresponding attention score. Patches with higher attention scores are deemed more important for the final prediction.

3.5.6 Conclusion

This paper introduces a novel text-guided representation learning pipeline designed for the efficient processing of Whole-Slide Images (WSIs). Our proposed model, mTREE, seamlessly integrates textual pathology information with WSI features on multiple levels, enabling a comprehensive understanding of the underlying data. Trained in an end-to-end manner, mTREE demonstrates superior performance in both classification and survival prediction across two distinct WSI datasets. Notably, the model exhibits explainability, as evidenced by its capability to visualize attention maps at both the WSI level and specific patches with high diagnostic importance. This fusion of accuracy and interpretability underscores the effectiveness of mTREE in the domain of WSI analysis.

CHAPTER 4

Energy Efficient Representation Learning for Meta-optics

4.1 Digital Modeling on Large Kernel Metamaterial Neural Network

4.1.1 Introduction

Digital neural networks (DNN) are essential in modern computer vision tasks. The convolutional neural network (CNN) is arguably the most widely used AI approach for image classification (LeCun et al., 1989; Krizhevsky et al., 2017; Li et al., 2014), segmentation (Jha et al., 2020; Ronneberger et al., 2015), and detection (Chauhan et al., 2018; Redmon et al., 2016). Even for more recent Vision Transformer-based models, convolution is still an essential component for extracting local image features (Liu et al., 2021b; Wang et al., 2021b; Liu et al., 2022b; Ding et al., 2022; Liu et al., 2022a). Current CNNs are typically deployed with computational units (e.g., CPUs and GPUs). Such a design might lead to a heavy computational burden, significant latency, and intensive power consumption, which are critical limitations in applications such as the Internet of Things (IoT), edge computing, and the usage of drones. Therefore, the AI community has started to seek DNN models with less energy consumption and lower latency. However, we might never approach energy-free and light-speed DNN following the current trends in research.

Fortunately, the recent advances in optical computational units (e.g., metamaterial) have shed light on energy-free and light-speed neural networks (Fig. 4.1). At its current stage, the SOTA metamaterial neural network (MNN) is implemented as a hybrid system, where the optical processors are used as a light-speed and energy-free front-end convolutional operator with a digital feature aggregator. Such design reduces the computational latency since the convolution operations are implemented by optical units, which off-loads more than 90 percent of the floating-point operations (FLOPs) in conventional CNN backbones like VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016). However, the digital design of the MNN is fundamentally limited by its physical structures, namely (1) **the optic system can only take positive value**; (2) **non-linear computations are challenging for free-space optic devices at low light intensity**; (3) **the implementation of the optical convolution is restricted by limited kernel size, channel number, precision, noise, and bandwidth**. Furthermore, limitations also exist in the current optic fabrication process: 1) **only the first layer of a neural network can be fabricated**, and 2) **limited layer capacity and weight precision**. Therefore, the unique advantages of the MNNs (e.g., light-speed computation) are not fully explored via standard 3×3 convolution kernels. The large convolution kernel (greater than 3×3) provides the larger reception fields which plays essential roles in segmentation and classification tasks (Long et al., 2015; Peng et al., 2017; Wang et al., 2020). Compared with traditional small kernel convolution (Geirhos et al., 2018), A larger receptive field (achieved using larger kernels or more convolutional layers) allows the network to see and model larger spatial contexts, which can be crucial in tasks where spatial details like boundaries matter (Cheng et al., 2020).

In this paper, we propose a novel large kernel metamaterial neural network (LMNN) that maximizes the digital capacity of the state-of-the-art (SOTA) MNN with model re-parametrization and network compression, while also considering the optical limitation explicitly. Our model maximizes the advantage of the light-speed natural of optical computing by implementing larger convolution kernels (e.g., 7×7, 11×11). The proposed LMNN yields larger reception fields, without sacrificing low computational latency and low energy consumption. Furthermore, the aforementioned physical limitations of LMNNs are explicitly addressed via optimized digital modeling. We evaluate our model on image classification tasks using two public datasets: FashionMNIST (Xiao et al., 2017) and STL-10 (Coates et al., 2011). The proposed LMNN achieved superior classification accuracy as compared with the SOTA MNN and model re-parametrization methods. Overall, the system's contributions can be summarized in four-fold:

- We propose the large convolution kernel design for an LMNN to achieve a larger reception field, lower computational latency, and less energy consumption.
- We introduce the model re-parameterization and multi-layer compression mechanism to compress the multi-layer multi-branch design to a single layer for the LMNN implementation. This maximizes the model capacity without introducing any extra burden during the optical inference stage.
- The physical limitations of LMNNs (e.g., limited kernel size, channel number, precision, noise, non-negative restriction, and bandwidth) are explicitly addressed via optimized digital modeling.

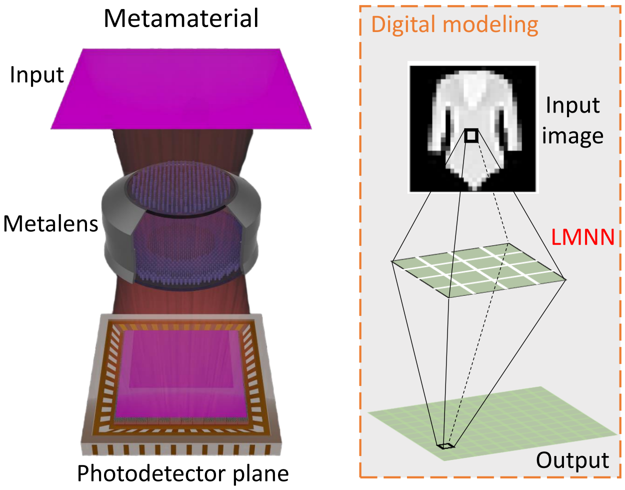


Figure 4.1: This study provides a digital modeling platform for designing and optimizing a metamaterial neural network (MNN). The proposed large kernel metamaterial neural network (LMNN) is able to maximize the performance of an MNN without introducing extra computational complexity during the inference stage.

- We implemented a one-layer LMNN with real physical metamaterial fabrication to demonstrate the feasibility of our hybrid design.

The rest of the paper is organized as follows. In Section II, we introduce background and related research relevant to large kernel convolution, re-parameterization, and optical neural networks. In Section III, our proposed LMNN model is presented. It includes the large kernel re-parameterization, meta-optic adaptation, and model compression strategy. Section IV focuses on presenting the dataset and experiment implementation details. Section V provides the experimental results and ablation study. Then, in Section VI and VII, we provide the discussion and conclude our work.

4.1.2 Related work

4.1.2.1 Models with large kernel convolution

For a decade, a common practice in choosing optimal kernel size in convolution is to leverage 3×3 kernel. In recent years, more attention has been put into a larger kernel design. The Inception network proposes an early design of adapting large kernels for vision recognition tasks (Chollet, 2017). After developing several variations (Szegedy et al., 2015, 2016), large kernel models became

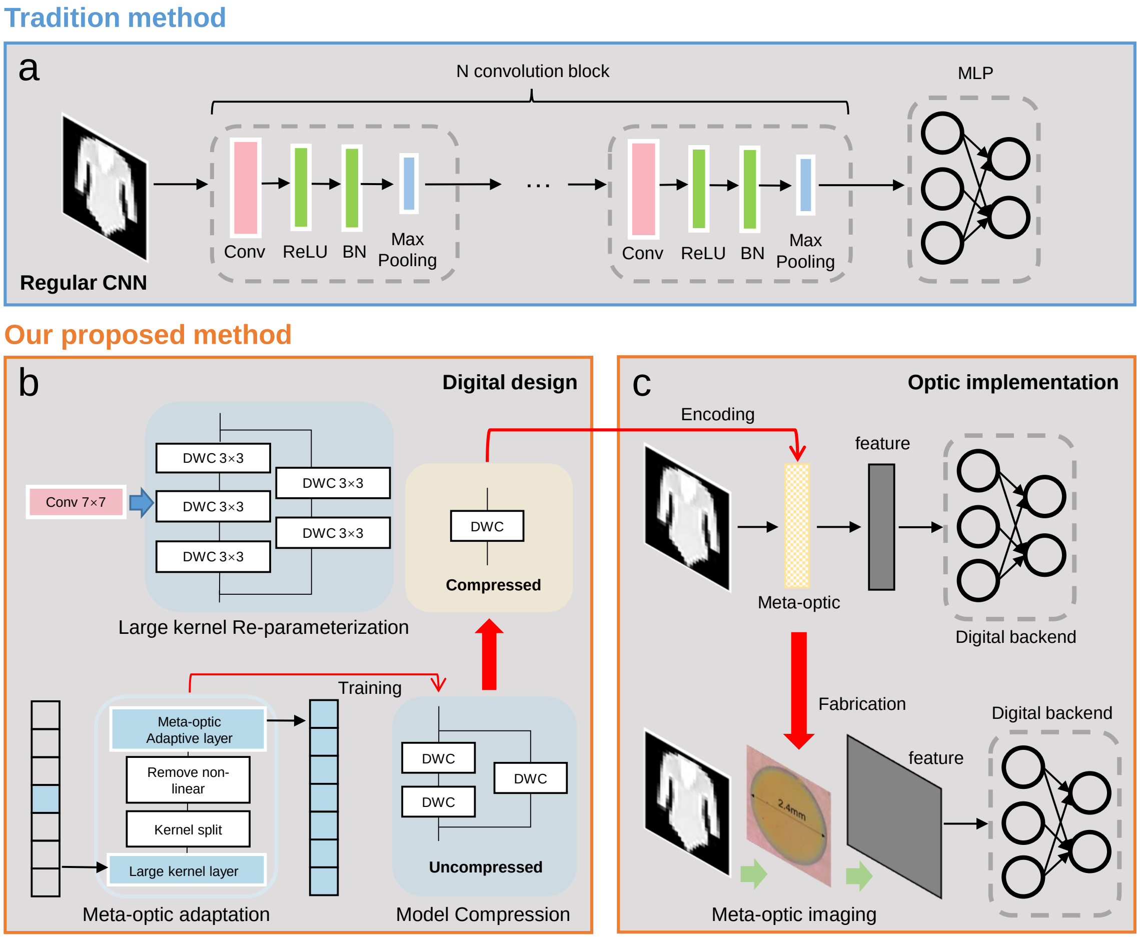


Figure 4.2: The upper panel (a) shows the conventional CNN model on the image classification task with Batch Normalization (BN) and Multilayer Perceptron (MLP). The lower panels (b) present our proposed LMNN method with digital design and optic implementation with depthwise convolution (DWC) layer. The large kernel re-parameterization efficiently achieves a large receptive field with a multi-branch multi-layer structure. Physical constraints are modeled via the meta-optic adaptation. The multi-branch multi-layer model is further compressed to a single-layer LMNN. (c) The digital design is fabricated as a real meta-optic device for inference. The red arrow shows the main pipeline to build the LMNN. The green arrow shows the image processing path in meta-optic imaging system.

less popular. Global Convolution Networks (GCNs) (Peng et al., 2017) employ the large kernel idea by utilizing 1×K followed by K×1 to achieve improvement in model performance for semantic segmentation.

Current limitations of leveraging large kernel convolution kernel can be divided into two aspects: (1) scaling up the kernel sizes leads to the degradation of model performance, and (2) its high computational complexity. According to the Local Relation Networks (LRNet) (Hu et al., 2019), the spatial aggregation mechanism with dynamic convolution is used to substitute traditional convolution operation. As compared with the traditional 3×3 kernels, the LRNet (Hu et al., 2019) leverages 7×7 convolution to improve model performance. However, the performance becomes saturated by scaling up the kernel size to 9×9. Similar to RepLKNet (Ding et al., 2022), scaling up the convolution kernel size to 31×31 without prior structural knowledge demonstrates the decrease of model performances. To leverage the heavyweight computation of large kernel convolution, (Sun et al., 2022) introduces the Shufflemixer for lightweight design.

4.1.2.2 Model compression and re-parameterization

Though many complicated ConvNets (Iandola et al., 2014; Huang et al., 2018) deliver higher accuracy than more simple ones, the drawbacks are significant. 1) The complicated multi-branch designs (e.g., residual addition in ResNet (He et al., 2016) and branch-concatenation in Inception (Szegedy et al., 2015)) make the model difficult to implement and customize, and slow down the inference and reduce memory utilization. 2) Some components (e.g., depthwise convolution in Xception (Chollet, 2017) and MobileNets (Howard et al., 2017), and channel shuffle in ShuffleNets (Zhang et al., 2018b)) increase memory access costs and lack support for various devices.

Model compression (Cheng et al., 2018) aims to reduce the model size and computational complexity (Vanhoucke et al., 2011; Chen et al., 2015) while maintaining their performance including pruning and quantization. Pruning has been widely used to compress deep learning models by removing the unnecessary or redundant parameters from a neural network without affecting its accuracy (Srinivas and Babu, 2015; Han et al., 2015; He et al., 2017). Quantization has two categories: Quantization-Aware Training (QAT) (Gong et al., 2014; Wu et al., 2016) and Post-Training Quantization (PTQ). QAT applies quantization operation in the training stage. In contrast, PTQ takes a full precision network for training and quantized it in the post stage (Liu et al., 2021c; Li et al., 2021).

Attempting to decrease the redundancy of CNN, SCConv (Li et al., 2023) compress the model by exploiting the spatial and channel redundancy among features.

4.1.2.3 Optical neural network

The Optical neural network uses light instead of electrical signals to perform matrix multiplications (Zhou and Anderson, 1994; Larger et al., 2012; Duport et al., 2012b) which can be much faster and more energy-efficient than traditional digital neural networks. Most optical neural networks (ONN) use a hybrid model structure: implement linear computation with optic device and non-linear operation digitally (Jutamulia and Yu, 1996; Paquot et al., 2012; Woods and Naughton, 2012; Hughes et al., 2018). Besides the use of optical devices, ONN has been implemented on nanophotonic circuits (Fang and Sun, 2015; Shen et al., 2017) and light-wave linear diffraction (Ovchinnikov et al., 1999; Lin et al., 2018) to improve model efficiency. For the non-linear computation, (George et al., 2018; Miscuglio et al., 2018) have proposed implementing the non-linear operation with the optic device on ONN.

4.1.3 Method

Problem statement. The goal of this study is to develop a new digital learning scheme to maximize the learning capacity of MNN while modeling the physical restrictions of meta-optic. With the proposed LMNN, the computation cost of the convolutional front-end can be offloaded into fabricated optical hardware, so as to get optimal energy and speed efficiency under current fabrication limitations. We adapt our innovations with four aspects: (1) large kernel re-parameterization, (2) meta-optic adaptation, and (3) model compression.

4.1.3.1 Large kernel re-parameterization

To tackle the limitation of only fabricating the first layer only in CNNs, we need to maximize the performance of the first layer, while it is feasible to adapt the fabrication processing. With the significant progress in Vision Transformers (ViTs), the key contribution for the performance gained is largely credited to the large effective receptive field, which can be generated similarly by the depthwise convolution with large kernel sizes in CNNs. Therefore, we explore the feasibility of adapting large kernel convolution in 1) single-branch and 2) multi-branch setting. The overarching

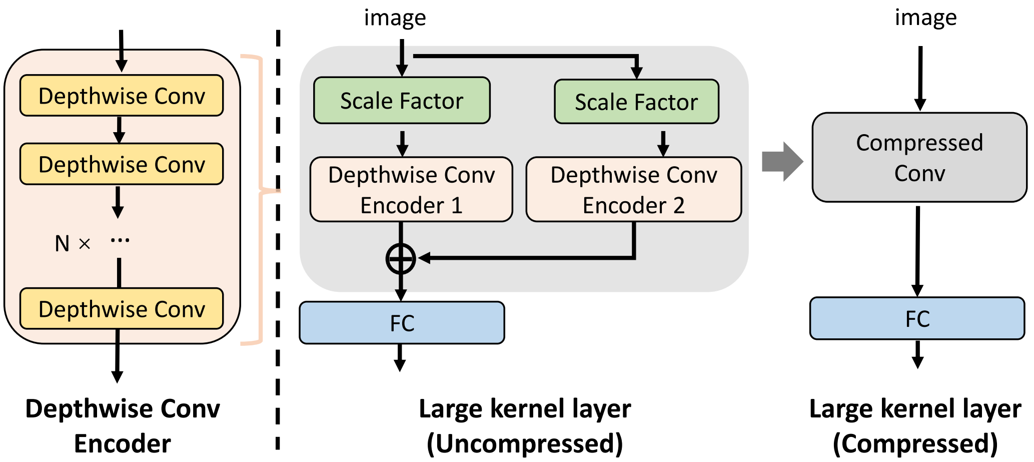


Figure 4.3: An overview of our proposed large convolution kernel block with re-parameterization is presented. Learnable scaling factors are employed to mimic the scaling function of batch normalization. In the inference stage, the block can be converted to a single convolution layer. ’FC’ refers to fully connected layer in the figure.

methodology for large kernel design can be delineated into two primary steps: (1) The deployment of stacked depthwise convolution layers to accommodate expansive convolution kernel receptive fields, as detailed in the ”Single Branch Design” section; (2) The amalgamation of results from various large convolution layers, each offering distinct scales of view, elaborated upon in the ”Multi-branch Design” section.

Single Branch Design. Inspired by (Ding et al., 2022), a large depthwise convolution kernel is equivalently to have the same receptive fields to a stack of small kernels. With the intrinsic structure of depthwise convolution, such a stack of kernel weights can be compressed into a single operator. It is thus essential for the LMNN to maximize the model performance via a relatively simple meta-optic design, with a single compressed convolution layer. The compressed design further introduces fewer model FLOPs in the model inference stage. For conventional convolution operation, the convolution weight matrix $\mathbf{W} \in \mathbb{R}^{C_i \times C_o \times K_h \times K_w}$. The C_i and C_o are input channel and output channel of the convolution layer. K_h and K_w are height and width of convolution kernel. Denote we have an input patch x in size of $H \times W$ and the output is y, we have conventional convolution as equation 4.1.

$$y = W * x \tag{4.1}$$

where $y = \sum_{p=0}^{n_i} W_p * x_p$, $*$ represents convolution between matrices. For the input x, the computation time complexity will be $O(H \times W \times C_i \times C_o \times K_h \times K_w)$. For the depthwise convolution model, channels C_i in the convolution layer are separated along with the input data channels of x. The depthwise convolution follows the equation. 4.2

$$y'_i = W_i * x_i \tag{4.2}$$

where y_i is the ith channel of output y, W_i and x_i are the ith channel from Convolution weight W and input data x, respectively. The time complexity is $O(H \times W \times C_i \times K_h \times K_w)$. Normally, the input channel number equals to the output channel number. We can infer the theoretical speed-up ratio r on model FLOPs between convention convolution and depthwise convolution following the equation.4.3

$$r = \frac{O(H \times W \times C_i \times C_o \times K_h \times K_w)}{O(H \times W \times C_i \times K_h \times K_w)} = O(C_i) \tag{4.3}$$

where C_i is the channel number of the convolution layer. Depthwise convolution has $C_i = C_o = C$. The depthwise convolution operation saves more FLOPs when the channel number is large compared with convention convolution.

Multi-branch design. Inspired by RepVGG (Ding et al., 2021) and RepLKNet (Ding et al., 2022), the multi-branch design demonstrates the feasibility of adapting large kernel convolutions (e.g., 31 × 31) with optimal convergence using a small kernel convolution in parallel. The addition of the encoder output enhances the large kernel convolution in the locality. According to the properties of convolution operation, the abstracted feature map from the parallel convolution path can be overlapped by learning different features. By using different convolution kernel sizes, the features from different scales of view are abstracted simultaneously.

We denote that output y' and input patch x use a two-branch convolution block W.

$$y'_i = W_1 * x + W_2 * x \tag{4.4}$$

where W_1 and W_2 is two different convolution layer with different kernel size. For multiple parallel paths, the N-branch convolution can be generated as equation 4.5.

$$y = \sum_{q=0}^{N} W_q * x \tag{4.5}$$

According to the equation 4.5, output y has the feature map from multiple scales of views. The overlap of convolution output from different scale redistributes the feature map which is proved by (Ding et al., 2022) to have better performance.

4.1.3.2 Meta-optic adaptation

As to integrate the large kernel convolution design into meta-optic devices, we need to consider and model the physical restrictions explicitly in our model design, beyond the conventional digital training (Fig. 4.2). First, the weight in convolution kernel should be positive for fabrication. Second, the convolution layer that substitutes by metalens should be the first layer of the model. Third, in this manuscript, metalens is designed at single wavelength (color). Thus, all RGB images are transferred to grayscale images. Fourth, due to the optic implementation purpose, the size of the convolution kernel is limited. Last, the channel number of the convolution layer is limited by the size of the optic device capacity.

Split kernel. To keep the model convolution kernel weight positive for the optic device implementation, we split the convolution kernel into two part: positive weight and negative weight. As shown in Fig.4.4, the final convolution kernel results are the subtraction of the two feature maps from the positive and negative convolution kernel respectively. Positively and negatively valued kernels are achieved for incoherent illumination by using polarization multiplexing, combined with a polarization-sensitive camera and optoelectronic subtraction.

Remove non-linear layer. In traditional convolution operation, non-linear layer is typically added between the convolution layers. The non-linear layer, including batch normalization and activation layers (eg. ReLU) introduce the non-linear transformation to the model. However, the nonlinear operation is not included in out meta-optic device due to the implementation cost. As shown in Fig.4.4, the non-linear layers are removed from the parallel convolution branch and connected behind the large kernel convolution layer.

Non-negative weight in optic kernel In traditional deep learning model, both positive weights and negative weights are stored. The meta-optic model implementation can only take positive kernel weights. Adaptation methods are applied to convolution model training to constrain model weight to a positive value. Four methods are introduced in model training: square of trigonometric functions, mask out the negative value, add non-negative loss, and our proposed kernel split. The former three methods constrain convolution kernel weight positive in digital model training. The last kernel split is achieved by meta-optic implementation.

Square of trigonometric function: instead of directly updating the weight, we define weight as equation 4.6. The weight W_i keeps positive and in range $[0,1]$ whatever the value of θ. To clarify, we utilize the square of the trigonometric function to constrain weights within the [0, 1] range during the model training process. This approach offers distinct advantages over normalizing the weights at the inference stage. Specifically, the parameter θ can be adjusted freely across any range without introducing negative weights, which is especially beneficial for our meta-optic implementation.

$$W_i = Sin^2(\theta_i) \tag{4.6}$$

Mask out the negative value: in the training process, the weight smaller than 0 is assigned as 0 manually after each iteration update.

Add non-negative loss: to maintain the model weight positive, a non-negative weight loss is added to the loss function, which is defined as equation 4.7.

$$loss = \sum(model.weight < 0) \tag{4.7}$$

Bandwidth and precision. Due to the accuracy of the current fabrication of meta-optic, the optical inference might lose precision. As a result, the model bandwidth and weight precision should also be modeled during the training process. For example, PyTorch has a default 32-bit precision, which is not feasible for the LMNN. Thus, the quantize is employed to simulate the model performance when all digital neural networks are implemented with optic devices. Taking the noise in optic implementation into consideration, which will affect the model weights precision, we add the Gaussian noise to the digital convolution weight.

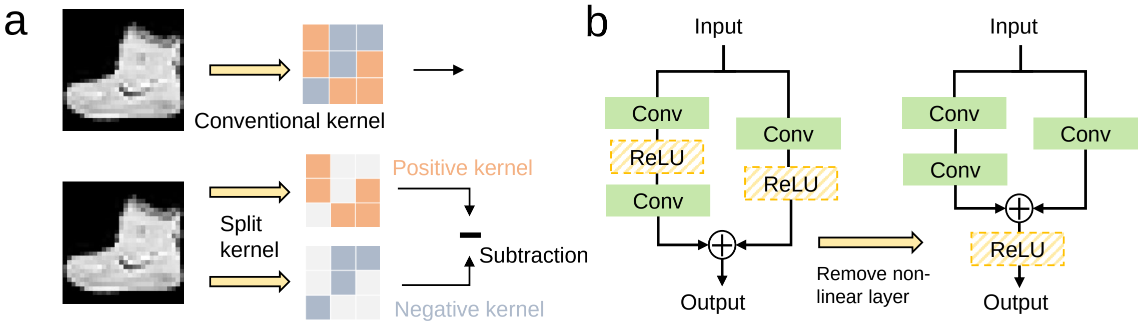


Figure 4.4: Adaptation for meta-optic implementation. (a) To implement the kernel with negative weight, we split the kernel into the positive kernel and negative kernel and subtract from their feature map. (b) The non-linear layer needs to be removed from the parallel convolution path.

4.1.3.3 Model compression

The stacked depthwise convolution and re-parameterization can potentially improve the model performance by learning with variance. The multilayer structure can be regarded as multiple stacked depthwise convolution layers which make the model deeper. The multi-branch structure will make the model wider. It is obvious the designed model is in a complex structure. To save image processing time in the inference stage, the multiplayer structure can be squeezed into a single layer. In this paper, we only explore the squeezed convolution layer. To get the equivalent squeezed layer, a non-linear component should be eliminated. The non-linear layers such as activation function and batch normalization are moved out of our squeezed block. The stacked convolution kernel follows the equation 4.8.

$$\begin{aligned} y =& (W_N * (W_{N-1} * \dots\dots\dots (W_2 * W_1))) * x \\ =& W^* * x \end{aligned} \tag{4.8}$$

$$W^* = (W_N * (W_{N-1} * \dots\dots\dots (W_2 * W_1))) \tag{4.9}$$

W^* is the equivalent weight to the stacked setting in equation 4.9. As the number of stacked convolution layers increases, the equivalent convolution kernel is larger. The equivalent kernel size k and the number of stacked 3×3 convolution layer n follow equation 4.10.

$$k = 2 \times n + 1 \tag{4.10}$$

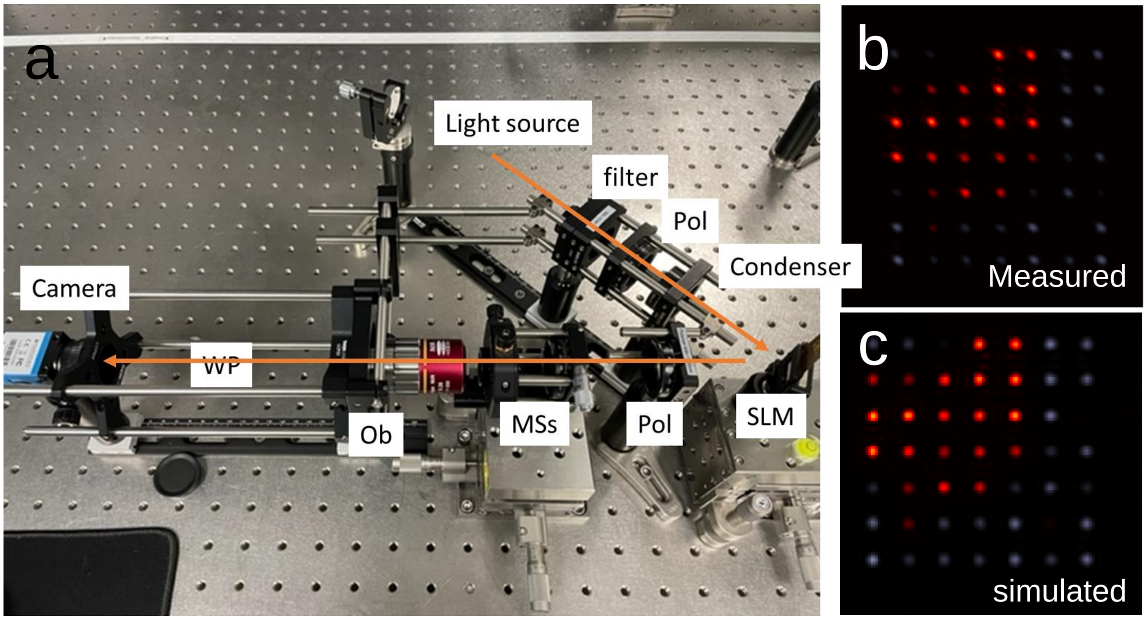


Figure 4.5: The meta-optic devices simulation and implementation platform. (a) Optic system for meta-optic lens test. The components in the figure are: Light source: Tungsten Lamp; filter: Wavelength filter; Pol: Polarizer; SLM: Spatial light modulator; Condenser: Lens to focus light on the SLM; MSs: Metasurfaces; Ob: Objective lens. (b) Measured meta-optic kernel weight point spread function, used for optical convolution with the imaged object. (c) Theoretical meta-optic kernel weight point spread function by simulation.

For example, two 3×3 convolution kernels are equivalent to a 5×5 convolution kernel. The multi-branch convolution layer can be compressed as shown in Fig 4.3.

Since the convolution kernel value from the different parallel branches is equivalent to a single kernel by overlapping kernel, a multi-parallel convolution branch can be compressed into a single path.

4.1.4 Data and experimental design

4.1.4.1 Data description

Two public datasets, FashionMNIST (Xiao et al., 2017) and STL10 (Coates et al., 2011), were employed to evaluate the performance of the proposed method on image classification tasks. For the FashionMNIST dataset, we employed 60,000 images for training and 10,000 images for testing. The images were grayscale images in the size of 28 × 28. FashionMNIST was inspired by the MNIST dataset, which classified clothing images rather than digits. We employed STL-10 as another cohort with a larger input image size (96×96). In our experiments, the RGB images in STL-10 were transferred to grayscale images due to the physical limitation in the LMNN.

4.1.4.2 Large kernel re-parameterization

We proposed the large re-parameterized convolution kernel design in our LMNN network to maximize the computational performance of the precious single metamaterial layer by (1) taking advantage of high-speed light computation, and (2) overcoming the physical limitations in an MNN implementation.

To evaluate the large re-parameterized convolution kernel on FashionMNIST, we constructed a naive model that consisted of a large re-parameterized convolution kernel block, a single fully connected layer, as well as non-linear components (ReLU activation, batch normalization, and the softmax function). Different re-parameterization model structures were evaluated. To demonstrate the impacts of the size, the kernel was tested from 3×3 to 31×31. Besides the kernel size, we evaluated multiple numbers of parallel branches: from a single path to four paths.

4.1.4.3 Meta-optic model adaptation

The performance of the LMNN is fundamentally limited by physical restrictions. We provide the model simulation by modeling optic system limitations. In regards to the model limitations, the convolution kernel is implemented with optical devices that can only have limited channels. To include the meta-optic devices in our network, the layer that is to be substituted should be the first layer of our model. The following model structure can be designed digitally. To validate model design on different sizes, deep neural networks with multiple convolution layers are implemented.

To simulate the noise in real meta-optic fabrication, we add random noise following the Gaussian distribution. To test the impact of noise level, we simulate the noise amplitude range from 0.05 to 0.2. Considering the meta-optic implementation on the whole model for further research, we quantize the model weight in 8-bit instead of the default 32-bit.

In order to evaluate the non-negative weight effect, three methods are evaluated to constrain the model weight positive. ’Sin’ in Fig. 4.8(a) means weights are defined by square of sin function. ’Mask out’ is to eliminate the negative weight by screening out. Loss function is also used to define model with positive weights, which results is shown in Fig. 4.8(a) as ’Non-neg’ loss.

The large kernel convolution design is validated on fabricated meta-optic devices. Based on the well-trained digital convolution kernel weight, meta-optic lenses are implemented and tested in real optic systems shown in Fig. 4.5. The imaging system using a liquid-crystal-based spatial light modulator (SLM) was built. An incoherent tungsten lamp with a bandpass filter was used for SLM illumination. The feature maps extracted by the meta-optic were recorded by a polarization-sensitive camera (DZK 33UX250, Imaging Source) where orthogonally polarized channels are simultaneously recorded using polarization filters on each camera pixel. The algorithm was programmed based on Pytorch 1.10.1 and CUDA 11.0 with a Quadro RTX 5000/PCIe/SSE2 as the graphics cards.

4.1.4.4 Model compression efficiency

Through model compression, the model in the inference stage alleviates the computation load with lighter weights. The fabricated convolution kernel by a meta-optic lens with the digital backend is assembled as the hybrid model. We test the model's inference time by feeding the same image and recording the model's processing time.

To test the optimal LMNN structure under the meta-optic fabrication limitation, the combination of layer numbers from one to five and channel numbers from nine to twenty. The model digital computation load (FLOPs) and the ratio of meta-optic is computed to find the model structure achieves optimal efficiency.

4.1.5 Result

In this section, we first evaluate our proposed large kernel network with a simple model structure, using the FashionMNIST dataset and STL-10 dataset. We then evaluate the large kernel capability on complex convolution neural networks with the same dataset.

4.1.5.1 Large re-parameterized convolution performance

We evaluate the large re-parameterized convolution model on two datasets: FashionMNIST and STL-10. As shown in Table. 4.1, the naive model with 7×7 convolution kernels has demonstrated better performance than that with 3×3. With structural reparameterization, the model prediction accuracy further improves. Meanwhile, the model implemented with a depthwise convolution layer

Table 4.1: Large re-parameterized convolution experiment results

	FashionMNIST		STL-10	
	Model Conv	Test	Model Conv	Test
Naive model	3×3	0.8495	3×3	0.4500
RepLKNet (Ding et al., 2022)	7×7	0.9015	7×7	0.4993
			11×11	0.5241
RepVGG (Ding et al., 2021)	7+5+3	0.9081	7+5+3	0.5341
			11+9+7	0.5650
Depthwise conv (Chollet, 2017)	3 dwc	0.9084	3 dwc	0.5509
			5 dwc	0.5935
Shufflemixer (Sun et al., 2022)	7×7	0.9047	7×7	0.5754
	11×11	0.9021	11×11	0.5878
SCConv (Li et al., 2023)	7×7	0.8975	7×7	0.5230
	11×11	0.8969	11×11	0.5117
LMNN (Ours)	3 dwc + 2 dwc + 1 dwc	**0.9115**	5 dwc + 3 dwc + 1 dwc	**0.6120**

’dwc’ refer to the depthwise convolution layer, convolution kernel size is 3 × 3

outperformed the baselines with both small and large convolution kernels. Other SOTA model performances are included: Shufflemixer reaches 0.5878 with 7×7 kernel while SCConv performs better on 11×11 kernel (0.5230).

Our large kernel model is evaluated on the STL10 dataset with a larger image size (96×96). As compared with performance on FashionMNIST (image size 30×30), the large kernel convolution model reveals greater improvements, as shown in Table. 4.1. The model with 11×11 kernel size has better accuracy (0.5341) compared with that of using 3×3 and 7×7. By integrating the depthwise convolution design, the model performance boosts from 0.5241 to 0.5935. Shufflemixer and SCConv is evaluated on STL10 with kernel size 7×7 and 11×11 and shows comparable model accuracy. Shufflemixer gets 0.9047 on 7×7 and SCConv gets 0.8975 on 11×11 kernel size. Our proposed large kernel block outperforms all SOTA approaches and achieves the best accuracy of 0.6015 with teacher model supervised training.

To further validate our large kernel with depthwise convolution design, we conduct experiments on more sophisticated models by replacing all convolution layers with the large re-parameterized convolution layers. Briefly, WideResNet-101 is used as complex model backbone (Kabir et al., 2022). Model performance is shown in Fig. 4.6. By substituting the first convolution layer with a larger kernel size, the model performance improves from 0.94 to 0.96 with utilizing larger images (256×256 RGB).

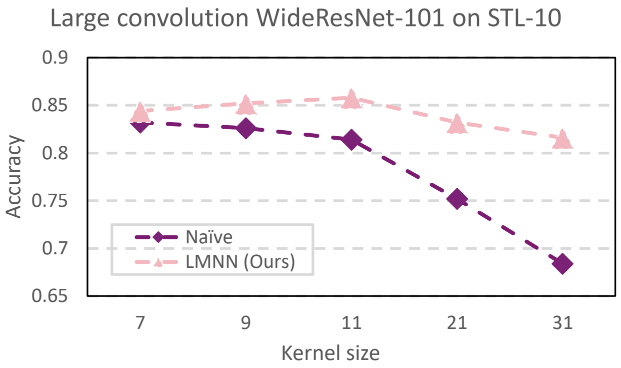


Figure 4.6: Performance of the Large Convolution WideResNet-101 model on the STL-10 dataset. The model was assessed using varying convolution kernel sizes, ranging from 7×7 to 31×31. The LMNN consistently outperforms, highlighting the benefits of utilizing larger convolution kernels.

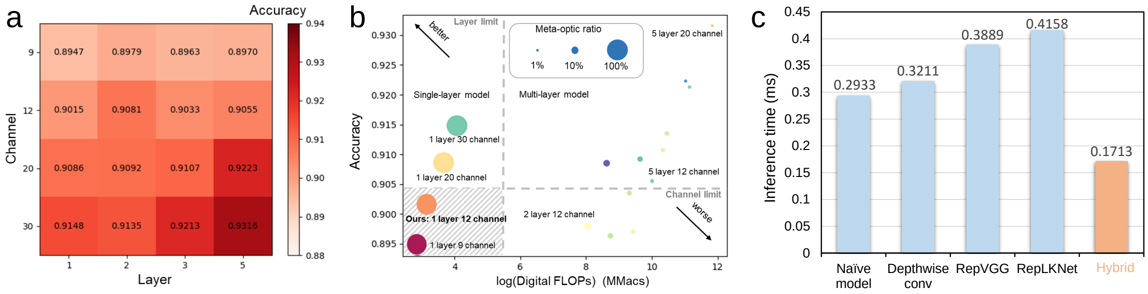


Figure 4.7: (a) The large re-parameterized convolution model performance with different layer numbers and channel numbers. (b) Large convolution kernel efficiency evaluation. The circle in different colors shows different convolution layer structures. The shadow area is the model structure that can be fabricated. The circle area shows the FLOPs ratio of the layer implemented by meta-optic material. x-axis is the model FLOPs except the layer to be fabricated. (c)Model inference time between the baseline digital model and hybrid model. The orange bar in the figure shows the time used our model.

4.1.5.2 Performance of model adaptation

To validate our large kernel design on the real metasurface fabrication model shown in Figure. 4.5, we implement a model trained on FashionMNIST with a large kernel design, utilizing a digital design for comparison. The digital convolution layer has 12 channels 7×7 convolution kernel which is the optimal kernel design under the current meta-optic implement limitation. As shown in Table.4.2, the Metamaterial Neural Network demonstrates excellent consistency with the theoretical performance of a digital neural network.

Due to the meta-optic implementation limitation, four adaptation methods are applied to constrain kernel weights to positive. According to the model performance, our proposed kernel split

Table 4.2: Metasurface fabrication

Method	Test
Digital Neural Network (DNN)	0.9015
Large Kernel MNN (LMNN)	0.8760

method shows superior performance over the common training strategies.

4.1.5.3 Ablation studies

To validate our model bandwidth and weight precision limit simulation, the results of the experiment are shown in Table 4.8.

To evaluate the upper bound performance on FashionMNIST, a deep model structure is implemented and tested on FashionMNIST. The number of convolution layers in our model ranges from 1 to 5, and the channel number ranges from 9 to 30. The model performance is shown in Fig. 4.7 (a). The model with more parameters shows a higher accuracy. Regardless of the meta-optic fabrication limitation, the meta-optic hybrid model achieves better performance.

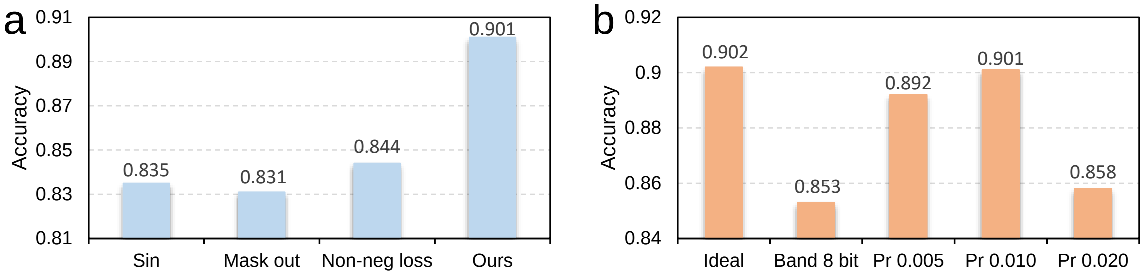


Figure 4.8: Plot of ablation study on LMNN. (a) Evaluating non-negative weight effect on model performance. (b) Measuring the effect of model bandwidth and weight precision effect on model prediction accuracy. 'Pr' in Figure (b) means precision.

4.1.5.4 LMNN efficiency and speed evaluation

To evaluate the model on both speed and computation load, we compute the model FLOPs except the large convolution layer and the FLOPs ratio of the layer implemented by meta-optic material. The model performance with different structure is shown in Figure. 4.7 (b). The optimal model structure should at top left corner in shadow area. As shown, the model with 1 large re-parameterized convolution layer and 12 channels is the optimal structure. To show the speed advantage of our LMNN, the model inference time is recorded. From Figure. 4.7 (c), the hybrid model shows a speed twice fast as compared to the digital convolution model.

4.1.6 Discussion

In this work, we present a convolution block with a large kernel design that generates larger receptive fields to maximize the digital capacity of LMNN. To validate the large kernel convolution design, we further applied the block to a complex model such as WideResNet-101. From the experiments, two important components contribute to the improvement of large kernel design from traditional 3×3 kernel size. First, the larger convolution kernel can get larger receptive fields. According to the target image size, the convolution kernel size is not the larger the better. For the FashionMNIST in size of 30×30, 7×7 is the best kernel size. For the images from STL-10 dataset in size of 96×96, 11×11 kernel performed the best. Another interesting point is the stacked depthwise convolution layers have equivalent computing operations to the single convolution layer with a larger kernel size. The multi-layer depthwise convolution and multi-branch structure expand the model capacity without parameter increase.

The proposed LMNN model bridges the disparity between natural objects and digital neural network analysis. Challenges in hybrid neural network design arise from the optical front-end, stemming from noise sources in the analog signals. These include stray light, detector interference, image misalignment due to optical inconsistencies, off-axis imaging aberrations, and fabrication flaws in the metalens and kernel layers. The system's bandwidth is constrained by the multi-channel lens, given the kernel layer's broadband nature. Optimizing the balance between bandwidth and aperture size is crucial for meta-optic systems. While the current optical approach mainly supports linear operations, future layers based on nonlinear media might facilitate activation functions. Even without these functions, refining the neural architecture can shift more linear tasks to the front-end. End-to-end model optimization ensures the meta-optic system effectively balances bandwidth and aperture considerations.

Since the large convolution kernel achieved superior performance on image classification task, more computer vision tasks have the improvement potential. For the image segmentation task, it can be regarded as a pixel-level classification problem. The large convolution design can be applied on segmentation tasks. Object detection can be another choice for large convolution kernel application. Different size of convolution kernel provides multiple field of views. The views from multiple scale can abstract representation with more spatial information.

4.1.7 Conclusion

In this study, we introduced a large-kernel convolution block tailored for implementation on a meta-optic lens. Through model re-parameterization and multi-layer compression, we were able to efficiently condense intricate digital layers, making them compatible with the constraints posed by optical fabrication techniques. By explicitly incorporating the physical restrictions, we re-evaluated and refined the design of a metamaterial neural network. The proposed LMNN demonstrated superior performance on the FashionMNIST and STL-10 datasets, attributable to its expanded receptive fields. Notably, the incorporation of light-speed optical convolution led to reductions in computational latency and energy consumption. Our research underscores the efficacy of optimized digital modeling, presenting a strategic pathway for adapting to physical limits in future optic-digital hybrid designs.

4.2 High-speed lightweight image segmentation by remodeling multi-channel meta-imagers

4.2.1 Introduction

In the realm of modern computer vision, digital neural networks (DNNs) play a pivotal role. The convolutional neural network (CNN) stands out as arguably the most extensively employed AI approach, particularly in tasks like image classification, segmentation, and detection. Despite the advent of Vision Transformer-based models, convolution remains integral for extracting local image features. Presently, CNNs are typically implemented on computational units like CPUs and GPUs. However, this conventional design approach brings forth substantial challenges, including a formidable computational load, notable latency issues, and heightened power consumption. These limitations become particularly pronounced in applications such as the Internet of Things (IoT), edge computing, and drone operations. Recognizing the critical need for DNN models with reduced energy consumption and lower latency, the AI community has embarked on a quest for more efficient solutions. Despite these efforts, achieving energy-free and light-speed DNNs within the current research trends seems to be an elusive goal.

Recent breakthroughs in optical computational units, including metamaterials (refer to Fig. 4.9), have brought to light the potential for neural networks that operate without energy consumption and at unprecedented speeds. The current cutting-edge metamaterial neural network (MNN) takes on a hybrid form, leveraging optical processors as a light-speed and energy-free front-end convolutional operator alongside a digital feature aggregator. This inventive approach serves to significantly reduce computational latency. By assigning the convolution operations to optical units, more than 90 percent of the floating-point operations (FLOPs) inherent in conventional CNN backbones like VGG and ResNet are effectively off-loaded. This marks a noteworthy departure from traditional architectures, opening up new avenues for efficient and high-performance neural network designs. However, the hybrid design is fundamentally influenced by the physical structure including the limited kernel size, and channel number. Besides that, the hybrid system limited the optic fabrication on the first layer of the neural network.

Based on our proposed LMNN model, the hybrid design achieves promising performance on the classification task. However, there are limitations of LMNN, namely: (1) the LMNN model can only perform image classification tasks instead of model complex tasks like image segmentation and object detection; (2) input images are in low resolution (28×28), and (3) besides leverage the

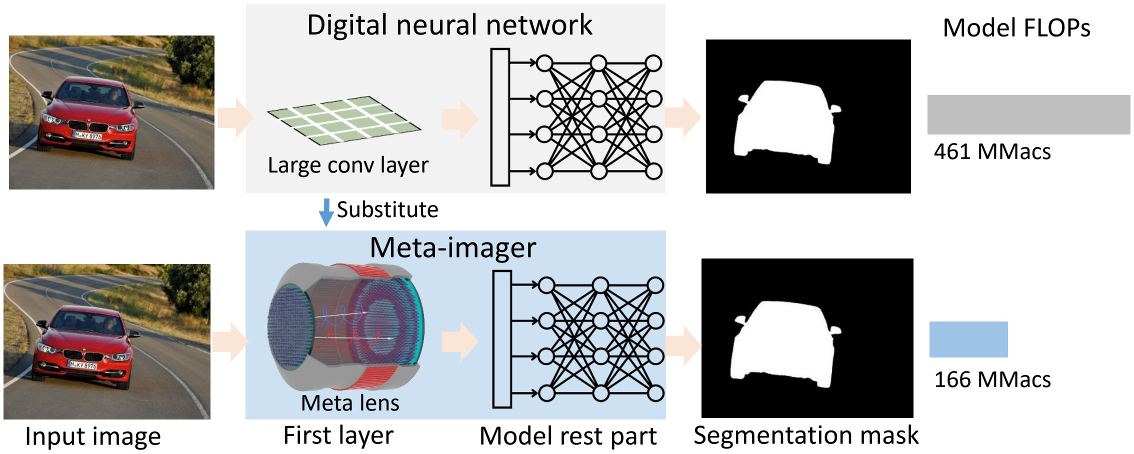


Figure 4.9: This study provides a hybrid pipeline for designing and optimizing a large kernel metamaterial neural network (MNN). The proposed Meta-imager is efficient for segmentation tasks with fewer FLOPs in computation.

computation burden to the optic, the digital part requires efficiency improvement operation like model compression in the inference stage.

In this paper, we propose a novel large kernel lightweight segmentation model Meta-imager that maximizes the efficiency advantages of optic signal computation, while also compressing the digital processing model to further improve the model segmentation efficiency. To adapt the segmentation task on large images, the proposed lightweight large kernel model achieves larger receptive fields, the ability to larger image analysis, and covers general vision tasks, image classification segmentation, and detection. Furthermore, the complexity of the model digital processing part is explicitly addressed via a set of model compression methods. We evaluate our design on image segmentation tasks using three public datasets: the portrait dataset, the Stanford dataset, and KITTI dataset. The proposed lightweight large kernel model achieved superior segmentation accuracy as compared with the SOTA segmentation model. Overall, the system's contributions can be summarized in four-fold:

- We propose a new large convolution kernel CNN network to achieve a large reception field, less energy consumption, and less latency.
- We introduce the model re-parameterization to improve large convolution kernel performance and sparse convolution kernel compression mechanism to compress the multi-branch sparse-convolution design to a single layer for the hybrid system implementation. The model com-

pression mechanism improves the model efficiency for digital processing.

- The task limitations of large convolution hybrid models are explicitly addressed via performing segmentation tasks on multiple datasets from different categories.

The rest of the paper is organized as follows. In Section II, we introduce background and related research relevant to large kernel convolution, model compression, and optical neural networks on image processing tasks. In Section III, our proposed lightweight lightspeed model is presented. It includes the large kernel re-parameterization, sparse convolution compression, and multi-path model compression. Section IV focuses on presenting the dataset and experiment implementation details. Section V provides the experimental results and ablation study. Then, in Sections VI and VII, we provide the discussion and conclude our work.

4.2.2 Related work

4.2.2.1 Large kernel convolution design

In the realm of convolutional neural networks (CNNs), the design and utilization of large kernel convolutions have garnered significant attention in recent years. Numerous studies have explored the benefits of using larger convolutional kernels, such as 7x7 or 11x11, to capture broader spatial contexts and more intricate patterns within images (Simonyan and Zisserman, 2014; Szegedy et al., 2015). Early research efforts focused on understanding the impact of kernel size on model performance, with findings suggesting that larger kernels can lead to improved feature extraction and recognition accuracy, especially for complex visual tasks (Zeiler and Fergus, 2014).

Building upon these findings, subsequent works have proposed various strategies to incorporate large kernel convolutions into CNN architectures effectively. These strategies often involve modifying network architectures, adjusting kernel sizes, or integrating multi-scale features to enhance the robustness and versatility of CNN models (Szegedy et al., 2016; He et al., 2016). Additionally, advancements in hardware acceleration and parallel processing have facilitated the efficient implementation of large kernel convolutions, enabling their widespread adoption across diverse computer vision applications (Zhang et al., 2018a; Sun et al., 2018).

Overall, the related work on large kernel convolution design underscores its pivotal role in advancing the capabilities of CNNs for tackling increasingly complex and demanding visual recognition tasks (Lin et al., 2013; Huang et al., 2017).

4.2.2.2 Optic neural network

Optic neural networks (ONNs) have emerged as a promising paradigm for accelerating neural network computations by leveraging the unique properties of optical computing. Inspired by the principles of light-based signal processing, ONNs exploit the parallelism, high bandwidth, and low energy consumption inherent in optical systems to achieve significant computational efficiency gains compared to traditional electronic implementations. A considerable body of research has focused on exploring various aspects of ONNs, including optical device design, system architectures, and algorithmic frameworks tailored to optical computing platforms (Shen et al., 2017; Lin et al., 2018; Hughes et al., 2018).

Early studies laid the groundwork for ONNs by demonstrating their potential for accelerating matrix-vector multiplications, a fundamental operation in neural network inference (Tait et al., 2017, 2016). Subsequent works have extended ONN capabilities to encompass more complex neural network layers and architectures, paving the way for practical applications in tasks such as image classification, object detection, and natural language processing (Miscuglio et al., 2018; Larger et al., 2012).

Key challenges in ONN research include addressing optical noise, device nonlinearity, and scalability issues, which require interdisciplinary efforts spanning optics, photonics, and machine learning (Jutamulia and Yu, 1996; Boehm et al., 2022). Despite these challenges, ONNs hold great promise for enabling ultra-fast and energy-efficient neural network computations, with the potential to revolutionize various domains of artificial intelligence and computing (Zhuge et al., 2021; Ovchinnikov et al., 1999).

4.2.2.3 Convolution neural network model compression

In the field of convolutional neural networks (CNNs), model compression techniques have garnered significant attention as a means to reduce the computational complexity and memory footprint of deep learning models without sacrificing performance. A diverse range of methods has been proposed to compress CNNs, including pruning, quantization, low-rank approximation, knowledge distillation, and weight sharing. Pruning techniques aim to remove redundant or less important parameters from the network, thereby reducing its size and computational cost (Han et al., 2015; Molchanov et al., 2016). Quantization methods reduce the precision of network parameters, often by representing weights and activations with fewer bits, to decrease memory requirements and improve inference speed (Hubara et al., 2017). Low-rank approximation techniques exploit the underlying structure of weight matrices to factorize them into smaller, more computationally efficient components (Denton et al., 2014). Knowledge distillation involves training a compact ”student” network to mimic the predictions of a larger ”teacher” network, transferring knowledge from the latter to the former (Hinton et al., 2015). Additionally, weight sharing approaches aim to reduce redundancy by sharing parameters across different parts of the network (Chen et al., 2015).

Collectively, these model compression techniques offer effective strategies for deploying CNNs on resource-constrained devices or accelerating inference in large-scale deployment scenarios. Ongoing research in this area continues to explore novel compression algorithms, optimization strategies, and application-specific considerations to further improve the efficiency and effectiveness of compressed CNN models.

4.2.3 Method

Problem statement We extensively study the trainability of large kernels on metamaterial neural networks (MNN) and unveil three main observations: (i) traditional convolution kernel shows limited improvement on large images; (ii) the MNN is only available on classification task; (iii) metamaterial implementation limited the computation ratio on segmentation model which is normally in complex structure.

4.2.3.1 Large convolution design with multiple path design

Limited by the image size and the task for the model, our previous proposed model LMNN achieved the prediction performance with kernel size 9×9. Two major limitations exist when applying the large kernel design to the MNN: (1) the metamaterial implementation limits the image size to a small range; (2) only the classification task is available to be validated on the MNN model when the segmentation task and detection task is too difficult to be implemented under the optic implementation limitation. To address the challenges, we proposed our model from two perspectives: (1) from kernel design, we employ the large convolution kernel with parameterization design to construct the

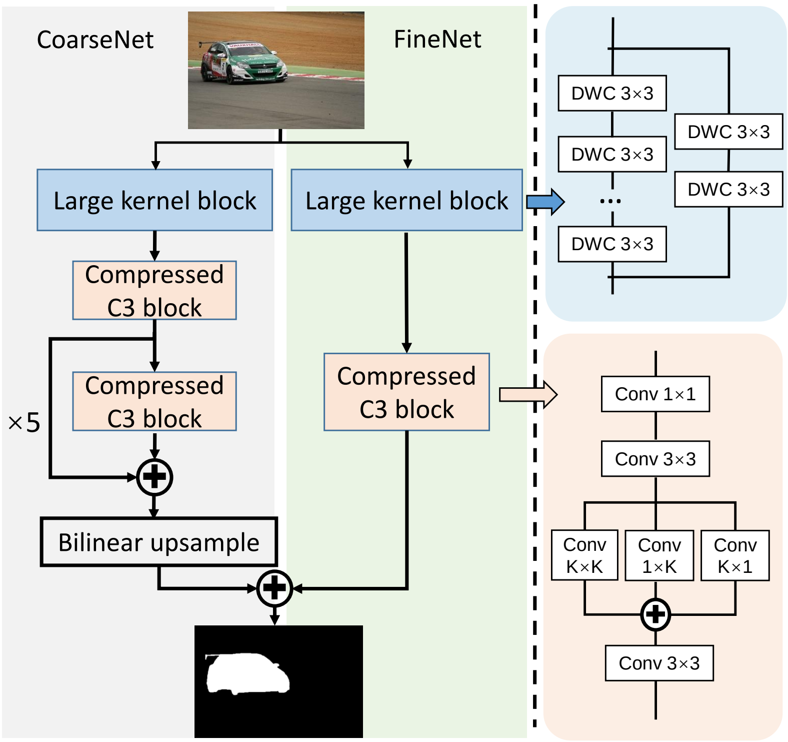


Figure 4.10: Lightweight segmentation model with hybrid meta optics design.

convolution layer (larger than 9×9); (2) from model design, our proposed lightweight segmentation model based on the multipath model structure composed by a course segmentation path and a light refinement path proposed by (Park et al., 2019).

4.2.3.2 Model compression with sparse convolution

Model compression is a crucial technique aimed at enhancing the efficiency of deep learning models by reducing their size and computational demands while maintaining their performance standards. Among the various strategies employed in model compression, pruning, and quantization stand out as widely adopted methodologies. Pruning, a prominent technique in model compression, involves the systematic removal of redundant or unnecessary parameters from neural networks. By identifying and eliminating connections that contribute minimally to the model's performance, pruning effectively reduces the model's size and computational requirements. This process allows for a more streamlined network architecture without sacrificing accuracy, making it particularly valuable for resource-constrained environments or deployment on edge devices.

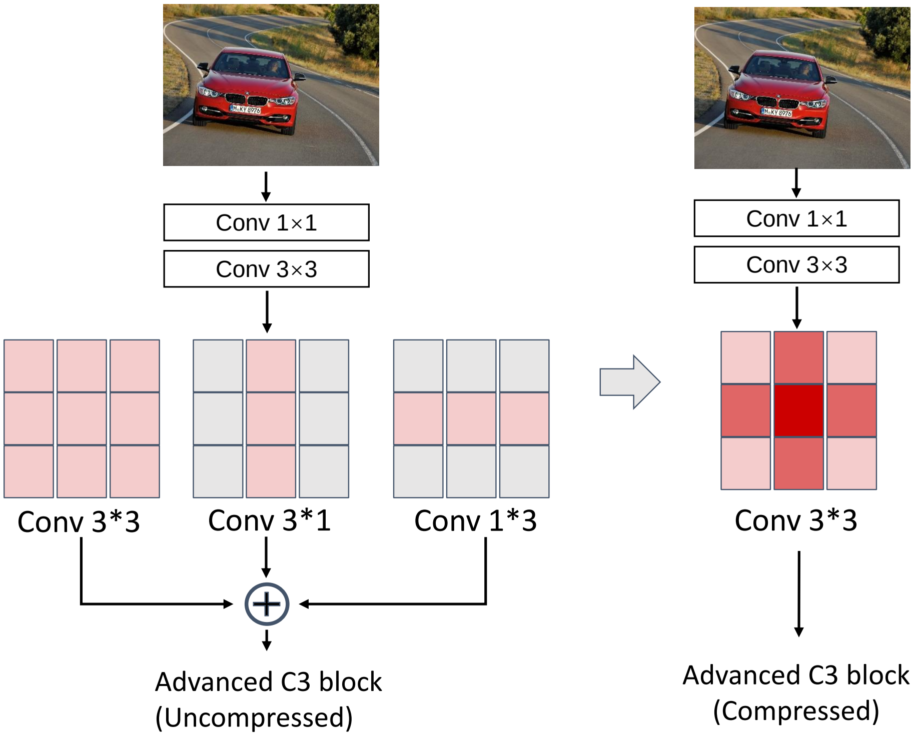


Figure 4.11: Model compression on segmentation model digital processing part.

We applied model compression and parameterization together for the sparse convolution kernel which is shown in Fig. Sparse convolution refers to a convolution operation where the kernel (filter) contains mostly zero values, resulting in a sparse structure. When using a kernel size of 1x3 (1 row and 3 columns), the convolution operation typically involves sliding this kernel over the input data and performing element-wise multiplication followed by summation along the spatial dimensions.

$$O_{h,w,c'} = \sum_{i=0}^{2} \sum_{j=0}^{C-1} I_{h,w+i,j} \times K_{0,i,j,c'} \tag{4.11}$$

I as the input tensor, K as the kernel tensor, O as the output tensor, and × as the convolution operation.

For the ExtremeC3 block, we have three convolution paths with kernel size $k \times k$, $1 \times k$, and $k \times 1$. The compressed convolution kernel follows Eq. Let's denote the individual kernels as $k_{1\times k}$, $k_{k\times k}$, $k_{k\times 1}$.

$$K_{combined}(i,j) = w_{1xk} \times K_{1xk}(i,j) + w_{kxk} \times K_{kxk}(i,j) + w_{kx1} \times K_{kx1}(i,j) \tag{4.12}$$

The compressed multipath convolution block saves computation complexity in the inference stage.

4.2.4 Data and experimental design

4.2.4.1 Data description

Three public datasets, EG1800 (Shen et al., 2016), Stanford Car dataset (Krause et al., 2013), and KITTI dataset (Geiger et al., 2012), were used to evaluate the lightweight large kernel model on segmentation tasks. For the EG1800 dataset, we employ 1887 images in 600 × 800 resolution with semantic segmentation masks. The EG1800 dataset is collected from Flicker with the manually annotated mask of the portrait. The Stanford Car dataset is composed of 16,185 RGB images of cars with the point coordinate where the car is located in images. The KITTI dataset is popular in mobile robotics and autonomous driving and features diverse traffic scenarios captured using high-resolution RGB, grayscale stereo cameras, and a 3D laser scanner. However, it lacks inherent ground truth annotations for semantic segmentation. To adapt to the segmentation task, both the Stanford Car dataset and the KITTI dataset need to address the annotation limitation.

4.2.4.2 Data generation with foundation model

Regarding the Stanford Car dataset and KITTI dataset lacking of segmentation annotation, we employ the Segment Anything Model (SAM) (Kirillov et al., 2023) to generate the object mask based on the prompts of object location. The SAM model is a foundation model that has a zero-shot ability to segment objects on new image distributions. The RGB image of Standford Car and KITTI datasets and bounding box coordinate is provided for the SAM model and SAM model will generate the object masks. With the help of the SAM model, the RGB images with object mask annotations are available for model training.

4.2.4.3 Large kernel digital design on segmentation model

The large kernel design is applied to the segmentation network's first convolution layer design. Since the first layer is designed to be substituted by the metaoptic lens in the inference stage, our large kernel design is under physical limitation. On the other hand, the optic lens provides light-speed computation which we can take advantage of. Based on the multipath segmentation network, the first convolution layers of both the Coarse-net part and Fine-net part are redesigned with the large convolution kernel with parameterization following the strategy in our previous work LMNN (Liu et al., 2023). Since the image is large compared with FashionMNIST previously used, our kernel size is larger from 9×9 to 15×15. The channel number is expanded from 12 to 48. The Larger convolution kernel and channel number provide the large capability of the first layers and handle the complex situation.

4.2.4.4 Model design with optic constrain

Under the meta-optic fabrication limitation, the meta-optic layer has limitations on both channel number and input size. The trade-off in model performance between input size and channel number is discussed. The size-first design uses the largest input image size under fabrication constrain. Channel-first design prefers more channel numbers under the fabrication limitation.

4.2.4.5 Model compression efficiency

Besides enlarging the capability of the first layer, our proposed lightweight segmentation network is compressed in the digital part. Since the model compression affects the model's complexity and efficiency, we evaluate if the compressed model loses accuracy. To test the efficiency of the model compression strategy, the model FLOPs, parameters, and FLOPs ratio of the first convolution layer.

4.2.5 Result

In this section, we evaluate our proposed lightweight segmentation network with a simple model structure, using the EG1800 dataset, Stanford Car dataset, and KITTI dataset. Since the Stanford Car dataset and KITTI dataset are car images, we train the model and test the two datasets together.

4.2.5.1 Segmentation performance on portrait dataset

We evaluate the lightweight segmentation model on EG1800 dataset together with model parameters and first convolution FLOPs ratio. As shown in Table. 4.3, the original ExtremeC3 model cannot take advantage of the large convolution kernel on the first layer, 15×15 kernel shows even lower performance than 11×11. The model performance without the first convolution layer shows a 2%

Table 4.3: Segmentation performance on EG1800

Model	Kernel size	1st conv FLOPs (%)	Model FLOPs	Digital FLOPs	Test (mIoU)
ExtremeC3	3×3	10.87	199.4	199.4	0.9249
	11×11	62.11	469.14	469.14	0.9323
	15×15	75.30	719.62	719.62	0.9301
Digital	N/A	N/A	174.10	174.10	0.9086
Ours	1×1	2.80	182.06	174.10	0.9137
	3×3	10.87	199.40	174.10	0.9234
	11×11	59.68	431.81	174.10	0.9415
	15×15	63.36	475.16	174.10	**0.9418**

Model FLOPs and digital FLOPs unit is MMAC.

Table 4.4: Segmentation performance on EG1800 after model compression

Model	Kernel size	1st conv FLOPs (%)	Model FLOPs	Digital FLOPs	Test (mIoU)
ExtremeC3	3×3	11.33	191.32	191.32	0.9233
	11×11	63.21	461.07	461.07	0.9315
	15×15	76.16	711.55	711.55	0.9289
Digital	N/A	N/A	166.03	166.03	0.9031
Ours	1×1	3.17	174.25	166.03	0.9121
	3×3	11.33	191.32	166.03	0.9217
	11×11	60.81	423.74	166.03	0.9404
	15×15	64.45	467.09	166.03	**0.9420**

Model FLOPs and digital FLOPs unit is MMAC.

drop compared with the ExtremeC3 model with 3×3 kernel size. Our proposed hybrid lightweight segmentation model achieves the best performance with 15×15 convolution kernel which has the same digital computation FLOPs.

Besides improving the model performance with advanced design on the first convolution layer, we evaluate the model efficiency improvement by model compression. Following the experiment setting in Table. 4.3, we applied model compression, including sparse convolution kernel compression and multipath parameterization, to each model design and shows the efficiency evaluation matrix in Table. 4.4. The compression method shows efficient computation on digital FLOPs without affecting model performance (mIoU).

4.2.5.2 Segmentation performance on car dataset

To validate our lightweight segmentation model with more datasets, we conduct experiments on the car dataset, including the Stanford Car dataset and KITTI dataset both with semantic segmentation mask as ground truth. Since the Stanford Car dataset and KITTI dataset are in different resolutions.

Table 4.5: Segmentation performance on car dataset

Model	Kernel size	Train (KITTI+Stanford)	Test (mIoU)	KITTI	Stanford
ExtremeC3	3*3	95.02	92.51	84.45	95.23
	11*11	95.12	92.09	84.37	95.39
	15*15	76.09	70.25	22.69	95.22
Digital	N/A	93.31	89.11	78.47	94.27
Ours	1*1	94.13	90.94	82.68	93.15
	3*3	94.97	92.01	85.05	94.77
	11*11	95.79	92.91	85.33	95.97
	15*15	96.05	93.17	87.41	95.19

Model FLOPs and digital FLOPs unit is MMAC.

Table 4.6: Segmentation performance on car dataset after model compression

Model	Kernel size	1st conv FLOPs (%)	Model FLOPs	Digital FLOPs	Test (mIoU)
ExtremeC3	3*3	11.33	191.32	191.32	91.36
	11*11	63.21	461.07	461.07	92.45
	15*15	76.16	711.55	711.55	70.01
Digital	N/A	N/A	166.03	166.03	88.97
Ours	1*1	3.17	174.25	166.03	90.94
	3*3	11.33	191.32	166.03	94.25
	11*11	60.81	423.74	166.03	95.32
	15*15	64.45	467.09	166.03	93.05

Model FLOPs and digital FLOPs unit is MMAC.

Both the Standford Car dataset and the KITTI dataset are used for model training.

4.2.5.3 Ablation studies

Due to the fabrication limitation of the meta-lens array, the priority of channel number and input image size need to be decided. The results of the experiment are shown in Figure 4.12. The left panel illustrates how increasing the input image size enhances performance compared to expanding the number of channels in a convolution layer. The gray area highlights the performance disparity in terms of mean Intersection over Union (mIoU). On the right panel, the effectiveness of utilizing large convolution kernels is assessed. Circles of various colors represent different convolution layer architectures, with the area of each circle indicating the ratio of FLOPs (Floating Point Operations per Second) for the layer when implemented using meta-optic materials. The x-axis represents the model's FLOPs, excluding the layer intended for fabrication.

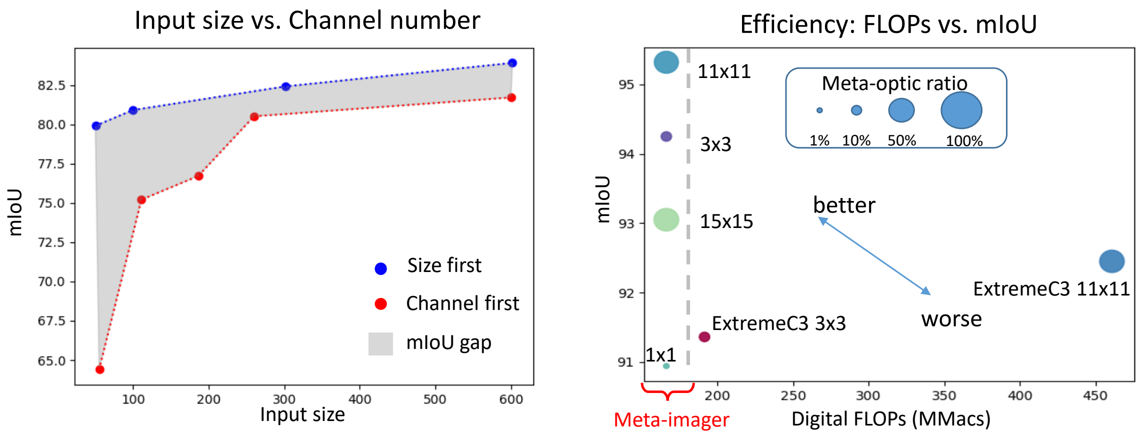


Figure 4.12: Model ablation study. Left panel: trade-off between input image size and channel number of convolution layer. Right panel: model efficiency visualization comparing model FLOPs and mIoU.

4.2.5.4 Model compression

Figure 4.13 demonstrates that the compressed model achieves a reduction of 8 MMacs in FLOPs, decreasing from 174.10 MMacs to 166.03 MMacs. The right panel indicates that the compressed model maintains equivalent performance to the original model. This consistency in performance illustrates that our meta-imager not only enhances the efficiency of the digital components but also contributes to the overall optimization of the hybrid system.

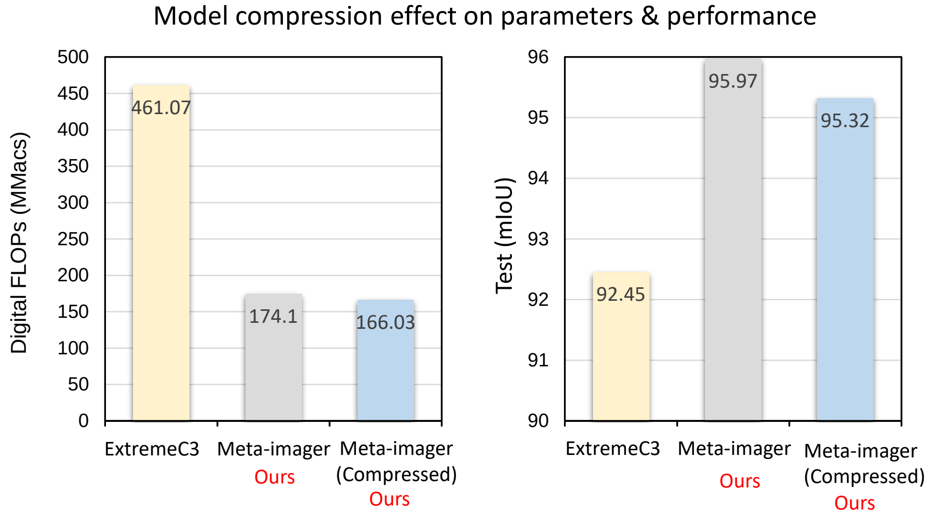


Figure 4.13: Model compression performance. Left panel: origin model, meta-imager, and compressed model parameters comparison; right panel: model performance after compression.

4.2.6 Discussion

Given the demonstrated superior performance of large convolution kernels in tasks such as image classification and segmentation, there exists substantial potential for their application in a wider array of complex computer vision tasks. Large convolution kernels have shown remarkable effectiveness in tasks like image classification and segmentation, primarily due to their ability to capture more extensive spatial information and intricate patterns within images. This success suggests that employing large convolution kernels in other computer vision tasks could yield significant improvements.

One such task is object detection, where accurately identifying and localizing objects within images is crucial. By utilizing large convolution kernels, the model can better discern the detailed features of objects, leading to more precise detection results. This can be particularly beneficial in scenarios with small or occluded objects, where finer details are essential for accurate recognition as the results shown in the experiments on car dataset.

Furthermore, in tasks involving image generation or synthesis, such as style transfer or super-resolution, large convolution kernels can enhance the model's ability to capture intricate textures and details, resulting in more realistic and high-fidelity output images. These kernels can effectively extract and preserve fine-grained features, which are instrumental in faithfully replicating the characteristics of the input images.

The application can be extended to video processing tasks like action recognition or video segmentation, large convolution kernels can enhance the model's capability to analyze temporal and spatial dependencies across frames. By incorporating information from a broader context, these kernels enable more robust understanding of dynamic scenes, leading to improved performance in tasks requiring temporal coherence and contextual understanding.

The adoption of large convolution kernels holds promise for advancing various complex computer vision tasks beyond traditional image classification and segmentation. Their ability to capture intricate details and spatial relationships makes them a valuable tool for enhancing the performance and capabilities of computer vision models across diverse applications.

4.2.7 Conclusion

In conclusion, we have introduced a novel large kernel lightweight segmentation model that harnesses the efficiency advantages of optical signal computation while integrating digital processing model compression techniques to further enhance segmentation efficiency. Our model offers larger receptive fields tailored for segmentation tasks on large images, extending its applicability to various vision tasks including image classification, segmentation, and detection. Through extensive evaluations on diverse datasets, including the portrait, Stanford, and KITTI datasets, our proposed approach has demonstrated superior segmentation accuracy compared to state-of-the-art models. Our contributions encompass the introduction of a novel large convolution kernel CNN network for larger reception fields, reduced energy consumption, and lower latency, alongside the introduction of model re-parameterization and sparse convolution kernel compression mechanisms to enhance model performance and efficiency in digital processing. By explicitly addressing task limitations and conducting segmentation tasks on multiple datasets from different categories, our work represents a significant step forward in the development of efficient and effective segmentation models for a wide range of computer vision applications.

CHAPTER 5

Contribution and Future Work

5.1 Contributions

5.1.1 Contribution on annotation-free semantic segmentation with microscopic image

In this paper, we develop a deep learning based unsupervised semantic segmentation method for sub-cellular microvilli segmentation using fluorescence microscopy. Meanwhile, we evaluate the performance of micro-level matching strategy, which is enabled by the multi-channel nature of fluorescence images. The contributions of this study are three-fold: (1) We propose the first deep learning based unsupervised sub-cellular microvilli segmentation method; (2) We propose the micro-level matching to ensure the roughly same number of objects across two modalities within each mini-batch, without introducing extra human annotation efforts; (3) Comprehensive analyses are provided to evaluate the outcomes of different augmentation strategies when generating the simulated masks for unsupervised microvilli segmentation.

5.1.2 Contribution on annotation-free synthetic instance segmentation and tracking for microscope video

Inspired by the recent generative adversarial network (GAN) based annotation-free image segmentation, we propose a novel annotation-free synthetic instance segmentation and tracking (ASIST) algorithm for analyzing microscope videos of sub-cellular microvilli. The contributions of this paper are three-fold: (1) a new annotation-free video analysis paradigm is proposed. (2) aggregating the embedding-based instance segmentation and tracking with annotation-free synthetic learning as a holistic framework; and (3) to the best of our knowledge, this is first study to investigate microvilli instance segmentation and tracking using embedding based deep learning. From the experimental results, the proposed annotation-free method achieved superior performance compared with supervised learning.

5.1.3 Contribution on simple triplet representation learning (SimTriplet) approach on pathological images

In this paper, we propose a simple triplet-based representation learning approach (SimTriplet), taking advantage of the multi-view nature of pathological images, with effective learning by using only a single GPU with 16GB memory. We present a triplet similarity loss to maximize the similarity between two augmentation views of the same image and between adjacent image patches. The contribution of this paper is three-fold:

(1) The proposed SimTriplet method takes advantage of the multi-view nature of medical images beyond self-augmentation.

(2) This method minimizes both intra-sample and inter-sample similarities from positive image pairs, without the needs of negative samples.

(3) The proposed method can be trained using a single GPU setting with 16GB memory, with batch size = 128 for 224×224 images, via mixed precision training.

5.1.4 Contribution on memory efficiency methods for self-supervised learning on pathological image analysis

In this work, we applied these memory-efficient approaches into a self-supervised framework. The contribution of this paper is three-fold: (1) We combined previously independent GPU memory-efficient methods with self-supervised learning framework; (2) Our experiments are to maximize the memory efficiency via limited computational resources (a single GPU); (3) The self-supervised learning framework with GPU memory-efficient method allows a single GPU to triple the batch size that typically requires three GPUs. From the experimental results, contrastive learning model with larger batch size leads to higher accuracy enabled by GPU memory-efficient method on single GPU.

5.1.5 Contribution on leverage the trained supervised and self-supervised models for pathological image survival analysis

In this paper, we present a simple and low-cost joint representation tuning (JRT) to aggregate task-agnostic vision representation (supervised ImageNet pretrained models) and pathological specific feature representation (self-supervised TCGA pretrained models) for downstream tasks. Our contribution is in three-fold: (1) we adapt and aggregate classification-based supervised and self-supervised representation to survival prediction via joint representation tuning, (2) comprehensive analyses on prevalent strategies of pretrained models are conducted, (3) the joint representation tuning provides a simple, yet computationally efficient, perspective to leverage large-scale pretrained models for both cancer diagnosis and prognosis. The proposed JRT method improved the c-index from 0.705 to 0.731 on the TCGA brain cancer survival dataset. The feature-direct JRT (f-JRT) method achieved 60x training speedup while maintaining 0.707 c-index score.

5.1.6 Contribution on multimodal multi-level fusion on Colorectal cancer microsatellite instability prediction

In this paper, we introduce a new and effective multi-modal fusion pipeline for MSI prediction by combining decision-level fusion and feature-level fusion following Bayesian rules. We also investigated different fusion strategies and found the proposed fusion scheme achieved better results than those methods. The contributions of this paper are: 1) This study generalizes an MSI prediction pipeline in CRC utilizing radiology-guided knowledge. 2) To the best of our knowledge, we are the first to exploit a multi-level fusion strategy for using multi-modal data for MSI prediction. 3) Extensive experimental results suggest the effectiveness of our Bayesian-based multimodal multi-level fusion. It can reduce the gap between pathology and radiology predictions and achieve more robust and accurate fusions than other feature-level or decision-level methods.

5.1.7 Contribution on multi-level text-guided representation end-to-end learning for whole slide image analysis

Our study introduces a text-guided representation learning method aimed at improving efficiency and extracting features from vital image regions, thereby eliminating the confusion caused by unnecessary image patches. The key contributions of our work are fourfold: (1) We present the first efficient visual-language model for gigapixel WSIs, operating in a seamless end-to-end fashion. (2) We utilize text information to optimize learning strategies across multiple levels. (3) Our pipeline is weakly supervised at the WSI level, eliminating the need for patch-level annotations from pathologists. (4) Our model offers explainability by providing visualizations at different levels, such as attention maps and significant patches.

5.1.8 Contribution on large convolution kernel design on meta-optics image classification

The system contributions can be summarized in four-fold:

(1) We propose the large convolution kernel design for an LMNN to achieve a larger reception field, lower computational latency, and less energy consumption.

(2) We introduce the model re-parameterization and multi-layer compression mechanism to compress the multi-layer multi-branch design to a single layer for the LMNN implementation. This maximizes the model capacity without introducing any extra burden during the optical inference stage.

(3) The physical limitations of LMNNs (e.g., limited kernel size, channel number, precision, noise, non-negative restriction, and bandwidth) are explicitly addressed via optimized digital modeling.

(4) We implemented a one-layer LMNN with real physical meta-material fabrication to demonstrate the feasibility of our hybrid design.

5.1.9 Contribution on large convolution kernel design on meta-optics image segmentation

The system's contributions can be summarized in four-fold:

(1) We propose a new large convolution kernel CNN network to achieve a large reception field, less energy consumption, and less latency.

(2) We introduce the model re-parameterization to improve large convolution kernel performance and sparse convolution kernel compression mechanism to compress the multi-branch sparse-convolution design to a single layer for the hybrid system implementation. The model compression mechanism improves the model efficiency for digital processing.

(3) The task limitations of large convolution hybrid models are explicitly addressed via performing segmentation tasks on multiple datasets from different categories.

5.2 Future work

5.2.1 Exploring multi-modalities assisted representation learning on pathology image

For the representation learning on pathology images, the information could be limited by a single modality. In the medical domain, information modalities are rich, including pathology images, radiology images, and clinical information. Cooperating with multi-modalities, the model performance can be improved by generating the information from multiple resources compared with solely pathology images. A number of research works have proved that multi-modality learning can benefit model training. However, the multiple-modality data of the same patient might be incomplete. The complete modality data collection is challenging for hospitals. Therefore, it is promising to find an optimal and efficient approach to abstract information from patients cooperating with incomplete modality data.

5.2.2 Exploring the foundation model ability on medical image analysis

The integration of Visual Language Models (VLM) has emerged as a cutting-edge approach, transforming the interpretation of large-scale pathological images. VLMs harness the combined strengths of Natural Language Processing (NLP) and Computer Vision, facilitating a holistic understanding of complex visual information by leveraging insights from diverse domains. Unlike fine-tuning, VLM relies on prompt prediction within a template, as demonstrated in CLIP and CoOp. These language models excel in knowledge acquisition and zero-shot learning.

Utilizing pre-trained language models and tailoring them to the specific challenges of Whole Slide Image (WSI), researchers have made significant advancements in capturing contextual relationships and hierarchical structures within pathology images. These models facilitate the extraction of meaningful features and semantic understanding, thereby enhancing the interpretability of WSIs for tasks such as image classification, tumor detection, and prognosis prediction.

5.2.3 Exploring the representation learning on super-resolution image

To perform the representation on super large images, obstacles exist on both computing resource consumptions and large enough reception fields for large images. The researchers have put up multi-instance learning (MIL) as a solution to adapt the gigapixel-level image to CNNs for natural images like ResNet50, which only focuses on a limited area. Inspired by the attention map, the model can learn the attention map from the global view of a high-resolution image. Attention sampling strategies, such as region-based methods and attention-guided sampling, have been proposed to enhance computational efficiency in tasks such as object detection. The attention-based methods on large images cost a massive of time in data preprocessing to make the data ready for retrieval.

Therefore, an efficient method for learning global representation from super-resolution images is promising.

5.2.4 Exploring the Large Vision Model (LVM) on medical image analysis

As proposed by (Bai et al., 2023), a novel sequential modeling approach employs vision-sentence fusion for Large Language Model (LLM) training, overcoming the limitation of paired linguistic data. This approach utilizes a large vision model capable of performing various tasks such as segmentation, classification, and denoising through pretraining with diverse prompts. In medical imaging, tasks like color deconvolution and super-resolution are specific to natural image data. Pretraining the large vision model on medical image tasks opens up the potential to address multiple tasks within a single model.

References

Al-Kofahi, Y., Zaltsman, A., Graves, R., Marshall, W., and Rusu, M. (2018). A deep learning-based algorithm for 2-d cell segmentation in microscopy images. *BMC bioinformatics*, 19(1):1–11.

Arbelle, A., Reyes, J., Chen, J.-Y., Lahav, G., and Raviv, T. R. (2018). A probabilistic approach to joint cell tracking and segmentation in high-throughput microscopy videos. *Medical image analysis*, 47:140–152.

Attia, M., Samih, Y., Elkahky, A., and Kallmeyer, L. (2018). Multilingual multi-class sentiment classification using convolutional neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., and Norouzi, M. (2021). Big self-supervised models advance medical image classification.

Baghli, I., Benazzouz, M., and Chikh, M. A. (2020). Plasma cell identification based on evidential segmentation and supervised learning. *International Journal of Biomedical Engineering and Technology*, 32(4).

Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A., Darrell, T., Malik, J., and Efros, A. A. (2023). Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*.

Ban, Z., Liu, J., and Cao, L. (2018). Superpixel segmentation using gaussian mixture model. *IEEE Transactions on Image Processing*, 27(8).

Bao, H., Dong, L., and Wei, F. (2021). Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Bar, Y., Diamant, I., Wolf, L., and Greenspan, H. (2015). Deep learning with non-medical training used for chest pathology identification. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 94140V. International Society for Optics and Photonics.

Bardes, A., Ponce, J., and LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8).

Boehm, K. M. et al. (2022). Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*, 22(2).

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

Braman, N. et al. (2021). Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12905 LNCS.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309.

Campanella, G., Silva, V. W. K., and Fuchs, T. J. (2018). Terabyte-scale deep multiple instance learning for classification and localization in pathology. *arXiv preprint arXiv:1805.06983*.

Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*.

Chauhan, R., Ghanshala, K. K., and Joshi, R. (2018). Convolutional neural network (cnn) for image detection and recognition. In *2018 first international conference on secure cyber computing and communication (ICSCCC)*, pages 278–282. IEEE.

Chen, C., Dou, Q., Chen, H., Qin, J., and Heng, P.-A. (2019). Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 865–872.

Chen, J. et al. (2021a). Actnn: Reducing training memory footprint via 2-bit activation compressed training. https://github.com/ucbrise/actnn.

Chen, R. J. et al. (2021b). Abstract po-002: Pan-cancer integrative histology-genomic analysis via interpretable multimodal deep learning. *Clinical Cancer Research*, 27(5_Supplement).

Chen, R. J. et al. (2022a). Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging*, 41(4).

Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F., Rodig, S. J., Lindeman, N. I., and Mahmood, F. (2020a). Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*.

Chen, R. J., Lu, M. Y., Weng, W.-H., Chen, T. Y., Williamson, D. F., Manz, T., Shady, M., and Mahmood, F. (2021c). Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025.

Chen, R. J., Lu, M. Y., Williamson, D. F., Chen, T. Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al. (2022b). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020b). A simple framework for contrastive learning of visual representations. In *37th International Conference on Machine Learning, ICML 2020*, volume PartF168147-3.

Chen, W., Wilson, J., Tyree, S., Weinberger, K., and Chen, Y. (2015). Compressing neural networks with the hashing trick. In *International conference on machine learning*, pages 2285–2294. PMLR.

Chen, X. and He, K. (2020). Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*.

Chen, X. and He, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Chen, Y.-H., Krishna, T., Emer, J. S., and Sze, V. (2017). Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE Journal of Solid-State Circuits*, 52(1).

Cheng, X., Wang, P., Guan, C., and Yang, R. (2020). Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10615–10622.

Cheng, Y., Liu, W., and Xing, W. (2021). Weighted feature fusion and attention mechanism for object detection. *Journal of Electronic Imaging*, 30(2):023015–023015.

Cheng, Y., Wang, D., Zhou, P., and Zhang, T. (2018). Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35(1):126–136.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Ciga, O., Xu, T., and Martel, A. L. (2021). Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, page 100198.

Ciga, O., Xu, T., and Martel, A. L. (2022). Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7.

Coates, A., Ng, A., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings.

Costa, P. et al. (2018). End-to-end adversarial retinal image synthesis. *IEEE Transactions on Medical Imaging*, 37(3).

Costa, P., Galdran, A., Meyer, M. I., Abràmoff, M. D., Niemeijer, M., Mendonça, A. M., and Campilho, A. (2017). Towards adversarial retinal image synthesis. *arXiv preprint arXiv:1701.08974*.

Cui, C., Liu, H., Liu, Q., Deng, R., Asad, Z., Wang, Y., Zhao, S., Yang, H., Landman, B. A., and Huo, Y. (2022). Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 626–635. Springer.

David, L. et al. (2019). Applications of deep-learning in exploiting large-scale and heterogeneous compound data in industrial pharmaceutical research. *Frontiers in Pharmacology*, 10.

de Hauwer, C., Darro, F., Camby, I., Kiss, R., van Ham, P., and Decaesteker, C. (1999). In vitro motility evaluation of aggregated cancer cells by means of automatic image processing. *Cytometry*, 36(1).

de Hauwer, C. et al. (1998). Gastrin inhibits motility, decreases cell death levels and increases proliferation in human glioblastoma cell lines. *Journal of Neurobiology*, 37(3).

De Marinis, L., Cococcioni, M., Castoldi, P., and Andriolli, N. (2019). Photonic neural networks: A survey. *IEEE Access*, 7.

del Hougne, P., Imani, M. F., van Diebold, A., Horstmeyer, R., and Smith, D. R. (2020). Learned integrated sensing pipeline: Reconfigurable metasurface transceivers as trainable physical layer in an artificial neural network. *Advanced Science*, 7(3).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2010). Imagenet: A large-scale hierarchical image database.

Denton, E. L., Zaremba, W., Bruna, J., LeCun, Y., and Fergus, R. (2014). Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems*, pages 1269–1277.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, X., Zhang, X., Han, J., and Ding, G. (2022). Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11963–11975.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., and Sun, J. (2021). Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742.

Dong, D., Fang, M.-J., Tang, L., Shan, X.-H., Gao, J.-B., Giganti, F., Wang, R.-P., Chen, X., Wang, X.-X., Palumbo, D., et al. (2020). Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. *Annals of oncology*, 31(7):912–920.

Dong, Y., Liu, Q., Du, B., and Zhang, L. (2022). Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. *IEEE Transactions on Image Processing*, 31:1559–1572.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Drozdzal, M., Chartrand, G., Vorontsov, E., Shakeri, M., Di Jorio, L., Tang, A., Romero, A., Bengio, Y., Pal, C., and Kadoury, S. (2018). Learning normalized inputs for iterative estimation in medical image segmentation. *Medical image analysis*, 44:1–13.

Dunn, K. W. et al. (2019). Deepsynth: Three-dimensional nuclear segmentation of biological images using neural networks trained with synthetic data. *Scientific reports*, 9(1).

Duport, F., Schneider, B., Smerieri, A., Haelterman, M., and Massar, S. (2012a). All-optical reservoir computing. *Opt Express*, 20(20).

Duport, F., Schneider, B., Smerieri, A., Haelterman, M., and Massar, S. (2012b). All-optical reservoir computing. *Optics express*, 20(20):22783–22795.

Echle, A., Rindtorff, N. T., Brinker, T. J., Luedde, T., Pearson, A. T., and Kather, J. N. (2021). Deep learning in cancer pathology: a new generation of clinical biomarkers. *British Journal of Cancer*, 124(4).

Edwards, H. and Storkey, A. (2016). Towards a neural statistician. *arXiv preprint arXiv:1606.02185*.

Fang, Y. and Sun, M. (2015). Nanoplasmonic waveguides: Towards applications in integrated nanophotonic circuits. *Light: Science and Applications*, 4.

Feng, L., Liu, Z., Li, C., Li, Z., Lou, X., Shao, L., Wang, Y., Huang, Y., Chen, H., Pang, X., et al. (2022). Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. *The Lancet Digital Health*, 4(1):e8–e17.

Fu, L., Yu, H., Li, X., Przybyla, C. P., and Wang, S. (2022). Deep learning for object detection in materials-science images: A tutorial. *IEEE Signal Processing Magazine*, 39(1).

Gadermayr, M., Gupta, L., Appel, V., Boor, P., Klinkhammer, B. M., and Merhof, D. (2019). Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: A study on kidney histology. *IEEE Transactions on Medical Imaging*, 38(10).

Gaszczak, A., Breckon, T. P., and Han, J. (2011). Real-time people and vehicle detection from uav imagery. In *Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, volume 7878.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

George, J., Amin, R., Mehrabian, A., Khurgin, J., El-Ghazawi, T., Prucnal, P. R., and Sorger, V. J. (2018). Electrooptic nonlinear activation functions for vector matrix multiplications in optical neural networks. In *Signal Processing in Photonic Communications*, pages SpW4G–3. Optica Publishing Group.

Gerlich, D., Mattes, J., and Eils, R. (2003). Quantitative motion analysis and visualization of cellular structures. *Methods*, 29(1).

Golia Pernicka, J. S., Gagniere, J., Chakraborty, J., Yamashita, R., Nardo, L., Creasy, J. M., Petkovska, I., Do, R. R., Bates, D. D., Paroder, V., et al. (2019). Radiomics-based prediction of microsatellite instability in colorectal cancer at initial computed tomography evaluation. *Abdominal Radiology*, 44:3755–3763.

Gong, Y., Liu, L., Yang, M., and Bourdev, L. (2014). Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Grill, J.-B. et al. (2020). Bootstrap your own latent a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 2020-December.

Hamerly, R., Bernstein, L., Sludds, A., Soljačić, M., and Englund, D. (2019). Large-scale optical neural networks based on photoelectric multiplication. *Physical Review X*, 9(2).

Han, S., Mao, H., and Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

Han, S., Pool, J., Tran, J., and Dally, W. (2015). Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.

Hassanin, M., Anwar, S., Radwan, I., Khan, F. S., and Mian, A. (2022). Visual attention methods in deep learning: An in-depth survey. *arXiv preprint arXiv:2204.07756*.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020a). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

He, K., Liu, X., Li, M., Li, X., Yang, H., and Zhang, H. (2020b). Noninvasive kras mutation estimation in colorectal cancer using a deep learning method based on ct imaging. *BMC medical imaging*, 20:1–9.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December.

He, Y., Zhang, X., and Sun, J. (2017). Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1389–1397.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. In *arXiv preprint arXiv:1503.02531*.

Hjelm, R. D. et al. (2019). Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019*.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Hoffer, E. and Ailon, N. (2015). Deep metric learning using triplet network. In *Lecture Notes in Computer Science*, volume 9370.

Hore, S. et al. (2015). Finding contours of hippocampus brain cell using microscopic image analysis. *Journal of Advanced Microscopy Research*, 10(2).

Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., and Saltz, J. H. (2016a). Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., and Saltz, J. H. (2016b). Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2424–2433.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, H., Zhang, Z., Xie, Z., and Lin, S. (2019). Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473.

Huang, G., Liu, S., Van der Maaten, L., and Weinberger, K. Q. (2018). Condensenet: An efficient densenet using learned group convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2752–2761.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J., and Zou, J. (2023). A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316.

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., and Bengio, Y. (2017). Quantized neural networks: Training neural networks with low precision weights and activations. In *arXiv preprint arXiv:1609.07061*.

Hughes, T. W., Minkov, M., Williamson, I. A. D., Shi, Y., and Fan, S. (2018). Training of photonic neural networks through in situ backpropagation and gradient measurement: supplementary material. *Optica*, Part F127-.

Huo, Y., Deng, R., Liu, Q., Fogo, A. B., and Yang, H. (2021). Ai applications in renal pathology. *Kidney International*, 99(6).

Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R. G., and Landman, B. A. (2018a). Adversarial synthesis learning enables segmentation without target modality ground truth. In *Proceedings-International Symposium on Biomedical Imaging*, volume 2018-April.

Huo, Y., Xu, Z., Moon, H., Bao, S., Assad, A., Moyo, T. K., Savona, M. R., Abramson, R. G., and Landman, B. A. (2018b). Synseg-net: Synthetic segmentation without target modality ground truth. *IEEE transactions on medical imaging*, 38(4):1016–1025.

Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., and Keutzer, K. (2014). Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*.

Ihle, S., Reichmuth, A. M., Girardin, S., Han, H., Stauffer, F., Bonnin, A., Stampanoni, M., Vörös, J., and Forró, C. (2019a). Udct: Unsupervised data to content transformation with histogram-matching cycle-consistent generative adversarial networks. *bioRxiv*, page 563734.

Ihle, S. J. et al. (2019b). Unsupervised data to content transformation with histogram-matching cycle-consistent generative adversarial networks. *Nature Machine Intelligence*, 1(10).

Ilse, M., Tomczak, J., and Welling, M. (2018). Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings-30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January.

Jain, V. et al. (2007). Supervised learning of image restoration with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*.

Jarrett, D., Yoon, J., and van der Schaar, M. (2019). Dynamic prediction in clinical survival analysis using temporal convolutional networks. *IEEE journal of biomedical and health informatics*, 24(2):424–436.

Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P., and Johansen, H. D. (2020). Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*, pages 558–564. IEEE.

Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S. N., Rosaen, K., and Vasudevan, R. (2016). Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*.

Julio, G., Merindano, M. D., Canals, M., and Ralló, M. (2008). Image processing techniques to quantify microprojections on outer corneal epithelial cells. *Journal of Anatomy*, 212(6).

Jutamulia, S. and Yu, F. T. S. (1996). Overview of hybrid optical neural networks. *Opt Laser Technol*, 28(2 SPEC. ISS.).

Kabir, H. D., Abdar, M., Khosravi, A., Jalali, S. M. J., Atiya, A. F., Nahavandi, S., and Srinivasan, D. (2022). Spinalnet: Deep neural network with gradual input. *IEEE Transactions on Artificial Intelligence*.

Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International journal of computer vision*, 1(4).

Katharopoulos, A. and Fleuret, F. (2019). Processing megapixel images with deep attention-sampling models. In *International Conference on Machine Learning*, pages 3282–3291. PMLR.

Kather, J. N. and Calderaro, J. (2020). Development of ai-based pathology biomarkers in gastrointestinal and liver cancer. *Nature Reviews Gastroenterology and Hepatology*, 17(10).

Kather, J. N. et al. (2019a). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med*, 25(7).

Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N. A., Ferber, D., et al. (2019b). Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730.

Kheterpal, N., Parvate, K., Wu, C., Kreidieh, A., Vinitsky, E., and Bayen, A. (2018). Flow: Deep reinforcement learning for control in sumo. *EPiC Series in Engineering*, 2:134–151.

Kieffer, B., Babaie, M., Kalra, S., and Tizhoosh, H. R. (2017). Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE.

Kieffer, B. et al. (2018). Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *Proceedings of the 7th International Conference on Image Processing Theory, Tools and Applications, IPTA 2017*.

Kim, Y. J. et al. (2021). Paip 2019: Liver cancer segmentation challenge. *Med Image Anal*, 67.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

Korfhage, N., Mühling, M., Ringshandl, S., Becker, A., Schmeck, B., and Freisleben, B. (2020). Detection and segmentation of morphologically complex eukaryotic cells in fluorescence microscopy images via feature pyramid fusion. *PLOS Computational Biology*, 16(9):e1008179.

Kornilov, A. S. and Safonov, I. v. (2018). An overview of watershed algorithm implementations in open source libraries. *Journal of Imaging*, 4(10).

Kraus, O. Z., Ba, J. L., and Frey, B. J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59.

Krause, J., Deng, J., Stark, M., and Fei-Fei, L. (2013). Collecting a large-scale dataset of fine-grained cars.

Krishna, M. M., Neelima, M., Harshali, M., and Rao, M. V. G. (2018). Image classification using deep learning. *International Journal of Engineering and Technology(UAE)*, 7.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun ACM*, 60(6).

Kulharia, V., Chandra, S., Agrawal, A., Torr, P., and Tyagi, A. (2020). Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *European Conference on Computer Vision*, pages 290–308. Springer.

Larger, L. et al. (2012). Photonic information processing beyond turing: an optoelectronic implementation of reservoir computing. *Opt Express*, 20(3).

Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., and Ng, A. Y. (2011). On optimization methods for deep learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Lee, K. et al. (2021). Deep learning of histopathology images at the single cell level. *Frontiers in Artificial Intelligence*, 4.

Leskó, M., Kato, Z., Nagy, A., Gombos, I., Torok, Z., Vigh Jr, L., and Vigh, L. (2010). Live cell segmentation in fluorescence microscopy via graph cut. In *2010 20th International Conference on Pattern Recognition*, pages 1485–1488. IEEE.

Li, H., Chen, G., Li, G., and Yu, Y. (2019). Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7274–7283.

Li, J., Wen, Y., and He, L. (2023). Scconv: Spatial and channel reconstruction convolution for feature redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6153–6162.

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., and Chen, M. (2014). Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pages 844–848. IEEE.

Li, R., Yao, J., Zhu, X., Li, Y., and Huang, J. (2018). Graph cnn for survival analysis on whole slide pathological images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11071 LNCS.

Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., and Gu, S. (2021). Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Lin, X. et al. (2018). All-optical machine learning using diffractive deep neural networks. *Science (1979)*, 361(6406).

Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, W. J. (2017). Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*.

Lipkova, J., Chen, R. J., Chen, B., Lu, M. Y., Barbieri, M., Shao, D., Vaidya, A. J., Chen, C., Zhuang, L., Williamson, D. F., et al. (2022). Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*, 40(10):1095–1110.

Liskowski, P. and Krawiec, K. (2016). Segmenting retinal blood vessels with deep neural networks. *IEEE Transactions on Medical Imaging*, 35(11).

Liu, Q. et al. (2021a). Simtriplet: Simple triplet representation learning with a single gpu. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12902 LNCS.

Liu, Q., Gaeta, I. M., Millis, B. A., Tyska, M. J., and Huo, Y. (2020). Gan based unsupervised segmentation: Should we match the exact number of objects. *arXiv preprint arXiv:2010.11438*.

Liu, Q., Zheng, H., Swartz, B. T., Asad, Z., Kravchenko, I., Valentine, J. G., Huo, Y., et al. (2023). Digital modeling on large kernel metamaterial neural network. *arXiv preprint arXiv:2307.11862*.

Liu, S., Chen, T., Chen, X., Chen, X., Xiao, Q., Wu, B., Pechenizkiy, M., Mocanu, D., and Wang, Z. (2022a). More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv preprint arXiv:2207.03620*.

Liu, S., Gibson, E., Grbic, S., Xu, Z., Setio, A. A. A., Yang, J., Georgescu, B., and Comaniciu, D. (2018). Decompose to manipulate: manipulable object synthesis in 3d medical images with structured image decomposition. *arXiv preprint arXiv:1812.01737*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022b). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986.

Liu, Z., Wang, Y., Han, K., Zhang, W., Ma, S., and Gao, W. (2021c). Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*, 34:28092–28103.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Loshchilov, I. and Hutter, F. (2017). Sgdr: Stochastic gradient descent with warm restarts.

Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Zhang, A., Le, L. P., et al. (2023). Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914*.

Lu, M. Y., Chen, T. Y., Williamson, D. F., Zhao, M., Shady, M., Lipkova, J., and Mahmood, F. (2021a). Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110.

Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. (2021b). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*, 5(6).

Lu, Y., Jha, A., Deng, R., and Huo, Y. (2022). Contrastive learning meets transfer learning: a case study in medical image analysis.

Lu, Y., Jha, A., and Huo, Y. (2021c). Contrastive learning meets transfer learning: A case study in medical image analysis. *arXiv preprint arXiv:2103.03166*.

Luo, P., Yu, F. R., Chen, J., Li, J., and Leung, V. C. (2021). A novel adaptive gradient compression scheme: Reducing the communication overhead for distributed deep learning in the internet of things. *IEEE Internet Things J*, 8(14).

Madabhushi, A. (2009). Digital pathology image analysis: opportunities and challenges. *Imaging Med*, 1(1).

Marini, N., Otálora, S., Müller, H., and Atzori, M. (2021). Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. *Med Image Anal*, 73.

Maška, M., Ulman, V., Svoboda, D., Matula, P., Matula, P., Ederra, C., Urbiola, A., España, T., Venkatesan, S., Balak, D. M., et al. (2014). A benchmark for comparison of cell tracking algorithms. *Bioinformatics*, 30(11):1609–1617.

Matula, P., Maška, M., Sorokin, D. V., Matula, P., Ortiz-de Solórzano, C., and Kozubek, M. (2015). Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. *PloS one*, 10(12):e0144959.

Meenderink, L. M. et al. (2019). Actin dynamics drive microvillar motility and clustering during brush border assembly. *Developmental Cell*, 50(5).

Mennel, L., Symonowicz, J., Wachter, S., Polyushkin, D. K., Molina-Mendoza, A. J., and Mueller, T. (2020). Ultrafast machine vision with 2d material neural network image sensors. *Nature*, 579(7797).

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al. (2017). Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. (2018). Mixed precision training.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. (2022). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7).

Miscuglio, M. et al. (2018). All-optical nonlinear activation function for photonic neural networks [invited]. *Opt Mater Express*, 8(12).

Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). Learning to focus: A machine learning approach to visual attention. *arXiv preprint arXiv:1409.1259*.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Vega, J. E. V., Brat, D. J., and Cooper, L. A. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979.

Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M., and Van Valen, D. (2019). Deep learning for cellular image analysis. *Nature methods*, pages 1–14.

Molchanov, P., Tyree, S., Karras, T., Aila, T., and Kautz, J. (2016). Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*.

Mormont, R., Geurts, P., and Marée, R. (2020). Multi-task pre-training of deep neural networks for digital pathology. *IEEE journal of biomedical and health informatics*, 25(2):412–421.

Mormont, R., Geurts, P., and Maree, R. (2021). Multi-task pre-training of deep neural networks for digital pathology. *IEEE J Biomed Health Inform*, 25(2).

Narang, S. et al. (2018). Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Neshatpour, K., Homayoun, H., and Sasan, A. (2019). Icnn: The iterative convolutional neural network. *ACM Transactions on Embedded Computing Systems*, 18(6).

Ng, J. Y.-H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702.

Nguyen, H. T., Worring, M., and van den Boomgaard, R. (2003). Watersnakes: Energy-driven watershed segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(3).

Nimura, Y., Hayashi, Y., Kitasaka, T., and Mori, K. (2015). Automated torso organ segmentation from 3d ct images using structured perceptron and dual decomposition. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414.

Noroozi, M. and Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9910 LNCS.

Otsu, N. (1979). Threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-9(1).

Ovchinnikov, Y. B., Müller, J., Doery, M., Vredenbregt, E., Helmerson, K., Rolston, S., and Phillips, W. (1999). Diffraction of a released bose-einstein condensate by a pulsed standing light wave. *Physical review letters*, 83(2):284.

Pal, S., Ebrahimi, E., Zulfiqar, A., Fu, Y., Zhang, V., Migacz, S., Nellans, D., and Gupta, P. (2019a). Optimizing multi-gpu parallelization strategies for deep learning training. *IEEE Micro*, 39(5):91–101.

Pal, S. et al. (2019b). Optimizing multi-gpu parallelization strategies for deep learning training. *IEEE Micro*, 39(5).

Pantanowitz, L., Valenstein, P. N., Evans, A. J., Kaplan, K. J., Pfeifer, J. D., Wilbur, D. C., Collins, L. C., and Colgan, T. J. (2011). Review of the current state of whole slide imaging in pathology. *Journal of pathology informatics*, 2(1):36.

Paquot, Y., Duport, F., Smerieri, A., Dambre, J., Schrauwen, B., Haelterman, M., and Massar, S. (2012). Optoelectronic reservoir computing. *Scientific reports*, 2(1):287.

Park, H., Sjösund, L. L., Yoo, Y., Bang, J., and Kwak, N. (2019). Extremec3net: Extreme lightweight portrait segmentation networks using advanced c3-modules. *arXiv preprint arXiv:1908.03093*.

Pauchard, Y. et al. (2016). Interactive graph-cut segmentation for fast creation of finite element models from clinical ct data for hip fracture prediction. *Computational Methods in Biomechanics and Biomedical Engineering*, 19(16).

Payer, C., Štern, D., Neff, T., Bischof, H., and Urschler, M. (2018). Instance segmentation and tracking with cosine embeddings and recurrent hourglass networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–11. Springer.

Pei, Q. et al. (2022a). Pre-treatment ct-based radiomics nomogram for predicting microsatellite instability status in colorectal cancer. *Eur Radiol*, 32(1).

Pei, Q., Yi, X., Chen, C., Pang, P., Fu, Y., Lei, G., Chen, C., Tan, F., Gong, G., Li, Q., et al. (2022b). Pre-treatment ct-based radiomics nomogram for predicting microsatellite instability status in colorectal cancer. *European Radiology*, 32:714–724.

Peikari, M., Salama, S., Nofech-Mozes, S., and Martel, A. L. (2018). A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific reports*, 8(1):1–13.

Peng, C., Zhang, X., Yu, G., Luo, G., and Sun, J. (2017). Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361.

Pinidiyaarachchi, A. and Wählby, C. (2005). Seeded watersheds for combined segmentation and tracking of cells. In *International Conference on Image Analysis and Processing*, pages 336–343. Springer.

Pratt, W. K. (2007). *Digital Image Processing: PIKS Scientific Inside*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ragothaman, S., Narasimhan, S., Basavaraj, M. G., and Dewar, R. (2016). Unsupervised segmentation of cervical cell images using gaussian mixture model. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 70–75.

Rai, T., Morisi, A., Bacci, B., Bacon, N., Thomas, S., La Ragione, R., Bober, M., and Wells, K. (2019). Can imagenet feature maps be applied to small histopathological datasets for the classification of breast cancer metastatic tissue in whole slide images? In *Medical Imaging 2019: Digital Pathology*, volume 10956, pages 191–200. SPIE.

Raju, A., Yao, J., Haq, M. M., Jonnagaddala, J., and Huang, J. (2020). Graph attention multi-instance learning for accurate colorectal cancer staging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–539. Springer.

Ray, N. and Acton, S. T. (2004). Motion gradient vector flow: An external force for tracking rolling leukocytes with shape and size constrained active contours. *IEEE Transactions on Medical Imaging*, 23(12).

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Refai, H., Li, L., Teague, T. K., and Naukam, R. (2003). Automatic count of hepatocytes in microscopic images. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 2, pages II–1101. IEEE.

Ridler, T. and Calvard, S. (1978). Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-8(8).

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Roth, H. R., Lu, L., Seff, A., Cherry, K. M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., and Summers, R. M. (2014). A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part I 17*, pages 520–527. Springer.

Sahin, I. H., Akce, M., Alese, O., Shaib, W., Lesinski, G. B., El-Rayes, B., and Wu, C. (2019). Immune checkpoint inhibitors for the treatment of msi-h/mmr-d colorectal cancer and a perspective on resistance mechanisms. *British journal of cancer*, 121(10):809–818.

Saraswat, M. and Arya, K. (2014). Automated microscopic image analysis for leukocytes identification: A survey. *Micron*, 65:20–33.

Sato, Y., Chen, J., Zoroofi, R. A., Harada, N., Tamura, S., and Shiga, T. (1997). Automatic extraction and measurement of leukocyte motion in microvessels using spatiotemporal image analysis. *IEEE Transactions on Biomedical Engineering*, 44(4).

Shao, Z., Wang, Y., Chen, Y., Bian, H., Liu, S., Wang, H., and Zhang, Y. (2023). Lnpl-mil: Learning from noisy pseudo labels for promoting multiple instance learning in whole slide image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21495–21505.

Shen, X., Hertzmann, A., Jia, J., Paris, S., Price, B., Shechtman, E., and Sachs, I. (2016). Automatic portrait segmentation for image stylization. In *Computer Graphics Forum*, volume 35, pages 93–102. Wiley Online Library.

Shen, Y., Harris, N. C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H., Englund, D., et al. (2017). Deep learning with coherent nanophotonic circuits. *Nature photonics*, 11(7):441–446.

Shi, X., Wei, D., Zhang, Y., Lu, D., Ning, M., Chen, J., Ma, K., and Zheng, Y. (2022). Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *European Conference on Computer Vision*, pages 151–168. Springer.

Sidaway, P. (2020). Msi-h: a truly agnostic biomarker? *Nature Reviews Clinical Oncology*, 17(2).

Simard, P. Y., Steinkraus, D., Platt, J. C., et al. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Icdar*, volume 3.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Srinivas, S. and Babu, R. V. (2015). Data-free parameter pruning for deep neural networks. *arXiv preprint arXiv:1507.06149*.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October.

Sun, L., Pan, J., and Tang, J. (2022). Shufflemixer: An efficient convnet for image super-resolution. *Advances in Neural Information Processing Systems*, 35:17314–17326.

Sun, M., Liu, Z., Wang, X., Qiao, W., and Lin, K. (2018). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Tait, A. N., Nahmias, M. A., Shastri, B. J., Prucnal, P. R., and Harris, J. S. (2016). Optical implementation of deep networks. *Applied Optics*, 55(36):A71–A82.

Tait, A. N., Nahmias, M. A., Shastri, B. J., Prucnal, P. R., and Harris, J. S. (2017). The physics of optical neural networks. *Applied Physics Reviews*, 4(2):021105.

Tang, B., Li, A., Li, B., and Wang, M. (2019). Capsurv: Capsule network for survival analysis with whole slide pathological images. *IEEE Access*, 7.

Tellez, D., Van Der Laak, J., and Ciompi, F. (2018). Gigapixel whole-slide image classification using unsupervised image compression and contrastive training. In *Conference on Medical Imaging with Deep Learning*, number Midl 2018.

Thongprayoon, C., Kaewput, W., Kovvuru, K., Hansrivijit, P., Kanduri, S. R., Bathini, T., Chewcharat, A., Leeaphorn, N., Gonzalez-Suarez, M. L., and Cheungpasitporn, W. (2020). Promises of big data and artificial intelligence in nephrology and transplantation.

Tian, Y., Krishnan, D., and Isola, P. (2019). Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015a). Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015b). The cancer genome atlas (tcga): An immeasurable source of knowledge. *Wspolczesna Onkologia*, 1A.

Ulman, V., Maška, M., Magnusson, K. E., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Matula, P., Svoboda, D., Radojevic, M., et al. (2017). An objective comparison of cell-tracking algorithms. *Nature methods*, 14(12):1141–1152.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L.-J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117.

Ushizima, D. et al. (2022). Deep learning for alzheimer's disease: Mapping large-scale histological tau protein for neuroimaging biomarker validation. *Neuroimage*, 248.

Van Valen, D. A., Kudo, T., Lane, K. M., Macklin, D. N., Quach, N. T., DeFelice, M. M., Maayan, I., Tanouchi, Y., Ashley, E. A., and Covert, M. W. (2016). Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLoS computational biology*, 12(11):e1005177.

Vanhoucke, V., Senior, A., and Mao, M. Z. (2011). Improving the speed of neural networks on cpus. In *Deep Learning and Unsupervised Feature Learning Workshop, NIPS 2011*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, C., Xu, R., Lv, K., Xu, S., Meng, W., Zhang, Y., Fan, B., and Zhang, X. (2023). Attention weighted local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, C.-W. et al. (2022a). A weakly supervised deep learning method for guiding ovarian cancer treatment and identifying an effective biomarker. *Cancers (Basel)*, 14(7).

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., and Tang, X. (2017). Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Wang, H., Hu, X., Zhao, X., and Zhang, Y. (2021a). Wide weighted attention multi-scale network for accurate mr image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):962–975.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364.

Wang, S., Shi, J., Ye, Z., Dong, D., Yu, D., Zhou, M., Liu, Y., Gevaert, O., Wang, K., Zhu, Y., et al. (2019). Predicting egfr mutation status in lung adenocarcinoma on computed tomography image using deep learning. *European Respiratory Journal*, 53(3).

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. (2021b). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578.

Wang, X. et al. (2021c). Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12908 LNCS.

Wang, Y., Zhang, X., Xu, J., Sun, X., Zhao, X., Li, H., Liu, Y., Tian, J., Hao, X., Kong, X., et al. (2022b). The development of microscopic imaging technology and its application in micro-and nanotechnology. *Frontiers in Chemistry*, 10:931169.

Wei, J. W., Tafe, L. J., Linnik, Y. A., Vaickus, L. J., Tomita, N., and Hassanpour, S. (2019). Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific reports*, 9(1):3358.

Weinstein, B. G. (2018). A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3):533–545.

Woods, D. and Naughton, T. J. (2012). Photonic neural networks. *Nature Physics*, 8(4):257–259.

Wu, C., Yu, H., Lee, S., Peng, R., Takeuchi, I., and Li, M. (2021). Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network. *Nature Communications*, 12(1).

Wu, F., Liu, P., Fu, B., and Ye, F. (2022). Deepgcnmil: Multi-head attention guided multi-instance learning approach for whole-slide images survival analysis using graph convolutional networks. In *2022 14th International Conference on Machine Learning and Computing (ICMLC)*, pages 67–73.

Wu, J. et al. (2019). The value of single-source dual-energy ct imaging for discriminating microsatellite instability from microsatellite stability human colorectal cancer. *Eur Radiol*.

Wu, J., Leng, C., Wang, Y., Hu, Q., and Cheng, J. (2016). Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4820–4828.

Wu, K., Gauthier, D., and Levine, M. D. (1995). Live cell image segmentation. *IEEE Transactions on biomedical engineering*, 42(1):1–12.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015a). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Xu, Y., Jia, Z., Ai, Y., Zhang, F., Lai, M., and Chang, E.-C. (2015b). Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2015-August.

Xu, Y., Jia, Z., Wang, L.-B., Ai, Y., Zhang, F., Lai, M., Eric, I., and Chang, C. (2017). Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC bioinformatics*, 18(1):1–17.

Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., Eric, I., and Chang, C. (2014). Deep learning of feature representation with multiple instance learning for medical image analysis. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1626–1630. IEEE.

Xue, Z., Zhang, T., and Lin, L. (2022). Progress prediction of parkinson's disease based on graph wavelet transform and attention weighted random forest. *Expert Systems with Applications*, 203:117483.

Yamashita, R. et al. (2021). Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol*, 22(1).

Yang, P., Hong, Z., Yin, X., Zhu, C., and Jiang, R. (2021). Self-supervised visual representation learning for histopathological images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12902 LNCS.

Yao, J., Cao, K., Hou, Y., Zhou, J., Xia, Y., Nogues, I., Song, Q., Jiang, H., Ye, X., Lu, J., et al. (2023). Deep learning for fully automated prediction of overall survival in patients undergoing resection for pancreatic cancer: A retrospective multicenter study. *Annals of Surgery*, 278(1):e68–e79.

Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., and Huang, J. (2020). Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789.

Yu, H., Guo, D., Yan, Z., Liu, W., Simmons, J., Przybyla, C. P., and Wang, S. (2018). Unsupervised learning for large-scale fiber detection and tracking in microscopic material images. *arXiv preprint arXiv:1805.10256*.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017). Deep sets. *Advances in neural information processing systems*, 30.

Zamora, I., Lopez, N. G., Vilches, V. M., and Cordero, A. H. (2016). Extending the openai gym for robotics: a toolkit for reinforcement learning using ros and gazebo. *arXiv preprint arXiv:1608.05742*.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. *arXiv preprint arXiv:1311.2901*.

Zhang, H. et al. (2021). An optical neural chip for implementing complex-valued neural network. *Nature Communications*, 12(1).

Zhang, L., Lu, L., Nogues, I., Summers, R. M., Liu, S., and Yao, J. (2017). Deeppap: deep convolutional networks for cervical cell classification. *IEEE journal of biomedical and health informatics*, 21(6):1633–1643.

Zhang, Q., Wang, H., Lu, H., Won, D., and Yoon, S. W. (2018). Medical image synthesis with generative adversarial networks for tissue recognition. In *Proceedings-2018 IEEE*.

Zhang, Q., Wang, H., Lu, H., Won, D., and Yoon, S. W. (2018). Medical image synthesis with generative adversarial networks for tissue recognition. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 199–207.

Zhang, X., Zhou, X., Lin, M., and Sun, J. (2018a). Efficient and accurate approximations of nonlinear convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1984–1992.

Zhang, X., Zhou, X., Lin, M., and Sun, J. (2018b). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856.

Zhang, Z., Yang, L., and Zheng, Y. (2018c). 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*, pages 9242–9251.

Zhao, M. et al. (2021). Faster mean-shift: Gpu-accelerated clustering for cosine embedding-based cell segmentation and tracking. *Medical Image Analysis*, 71.

Zhao, M., Jha, A., Liu, Q., Millis, B. A., Mahadevan-Jansen, A., Lu, L., Landman, B. A., Tyskac, M. J., and Huo, Y. (2020a). Faster mean-shift: Gpu-accelerated embedding-clustering for cell segmentation and tracking. *arXiv preprint arXiv:2007.14283*.

Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., et al. (2020b). Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4837–4846.

Zheng, H., Fu, J., Zha, Z.-J., and Luo, J. (2019). Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5012–5021.

Zhou, G. and Anderson, D. Z. (1994). Acoustic signal recognition with a photorefractive time-delay neural network. *Opt Lett*, 19(9).

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Zhu, C. et al. (2017a). Retinal vessel segmentation in colour fundus images using extreme learning machine. *Computerized Medical Imaging and Graphics*, 55.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017b). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017c). Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.

Zhu, X., Yao, J., and Huang, J. (2016). Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547. IEEE.

Zhu, X., Yao, J., and Huang, J. (2017d). Deep convolutional neural network for survival analysis with pathological images. In *Proceedings-2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016*.

Zhuang, C., Zhai, A., and Yamins, D. (2019). Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October.

Zhuang, J. and Wang, D. (2020). Geometrically matched multi-source microscopic image synthesis using bidirectional adversarial networks. *arXiv preprint arXiv:2010.13308*.

Zhuang, J. and Wang, D. (2022). Geometrically matched multi-source microscopic image synthesis using bidirectional adversarial networks. In *Lecture Notes in Electrical Engineering*, volume 784 LNEE.

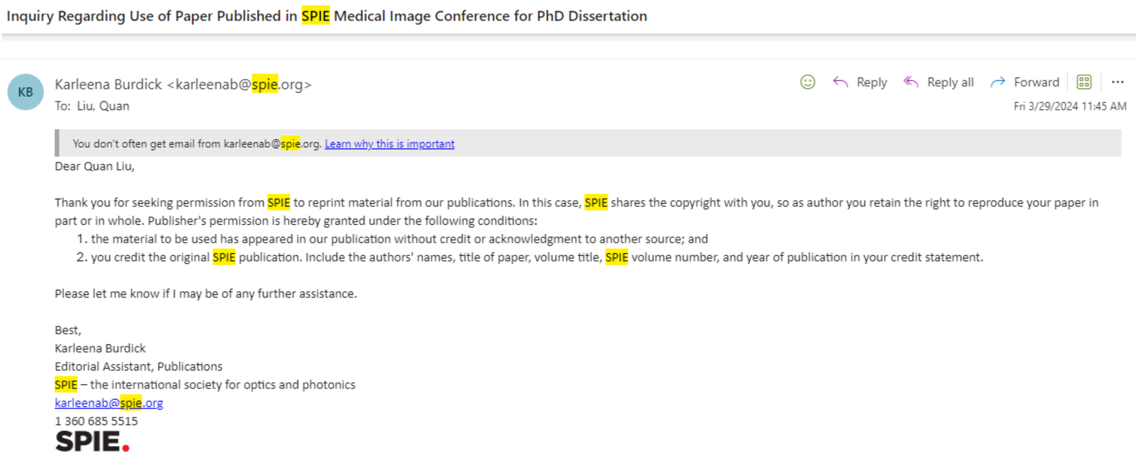
Zhuge, M. et al. (2021). Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Appendix A

Copyright from Publishers

A.1 Copyright from SPIE and JMI

This is the permission obtained from SPIE. The email screenshot is shown in A.1. Our content in Chapter 2.1, 3.2 is the publication in SPIE.



Inquiry Regarding Use of Paper Published in SPIE Medical Image Conference for PhD Dissertation

Karleena Burdick <karleenab@spie.org>
To: Liu, Quan

Fri 3/29/2024 11:45 AM

You don't often get email from karleenab@spie.org. Learn why this is important

Dear Quan Liu,

Thank you for seeking permission from SPIE to reprint material from our publications. In this case, SPIE shares the copyright with you, so as author you retain the right to reproduce your paper in part or in whole. Publisher's permission is hereby granted under the following conditions:

1. the material to be used has appeared in our publication without credit or acknowledgment to another source; and
2. you credit the original SPIE publication. Include the authors' names, title of paper, volume title, SPIE volume number, and year of publication in your credit statement.

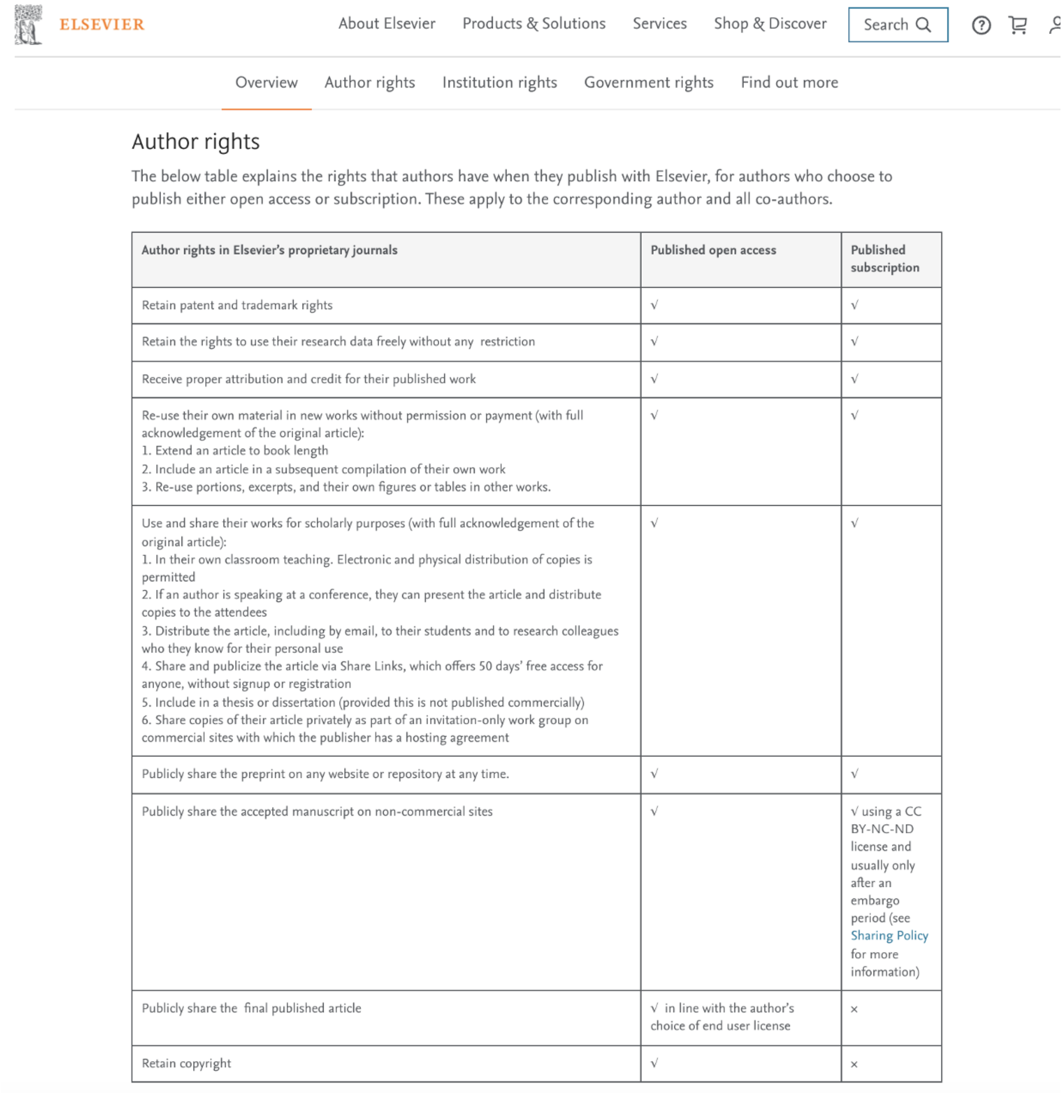
Please let me know if I may be of any further assistance.

Best,
Karleena Burdick
Editorial Assistant, Publications
SPIE – the international society for optics and photonics
karleenab@spie.org
1 360 685 5515
SPIE.

Figure A.1: Copyright from SPIE

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 Computers in Biology and Medicine paper (Chapter 2.2) is under Elsevier publisher.



Author rights

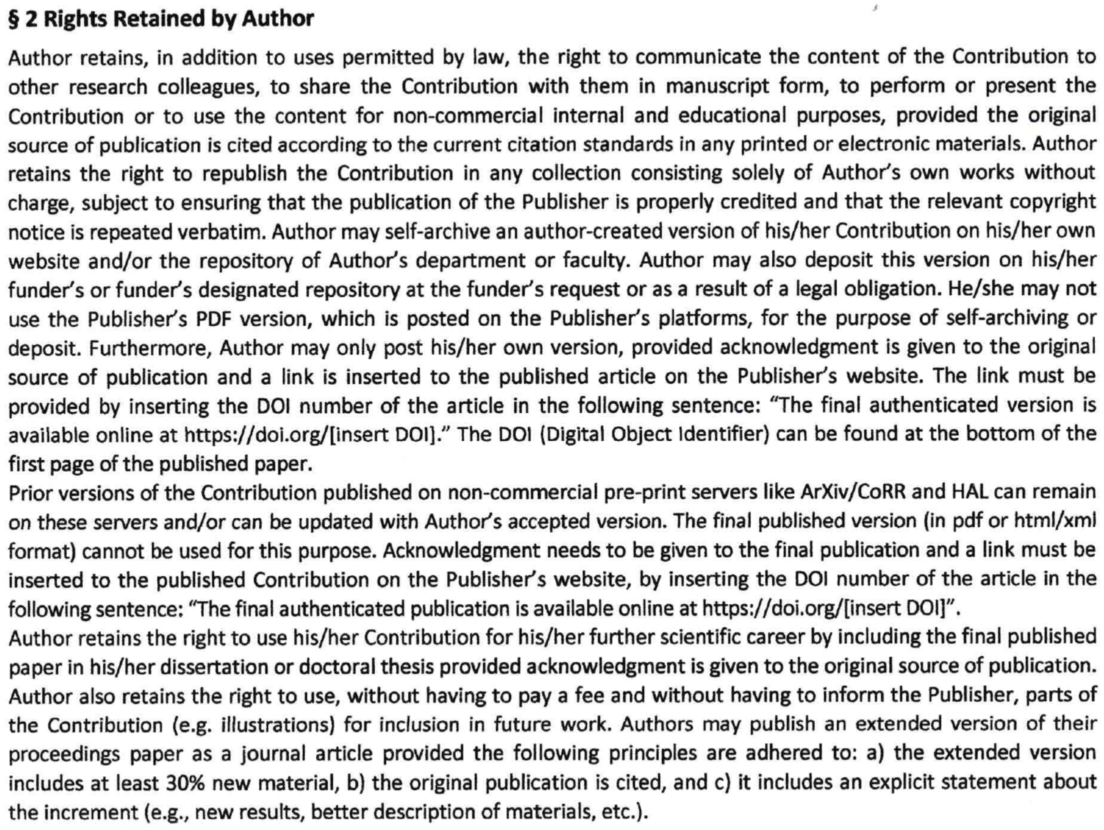
The below table explains the rights that authors have when they publish with Elsevier, for authors who choose to publish either open access or subscription. These apply to the corresponding author and all co-authors.

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	×
Retain copyright	√	×

Figure A.2: Copyright from Elsevier

A.3 Copyright from LNCS

Authors retains the right to use the content for mon-commercial internal and educational purposes, etc. Our Chapter 3.1 3.3 3.5 are 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]." The DOI (Digital Object Identifier) can be found at the bottom of the first page of the published paper.

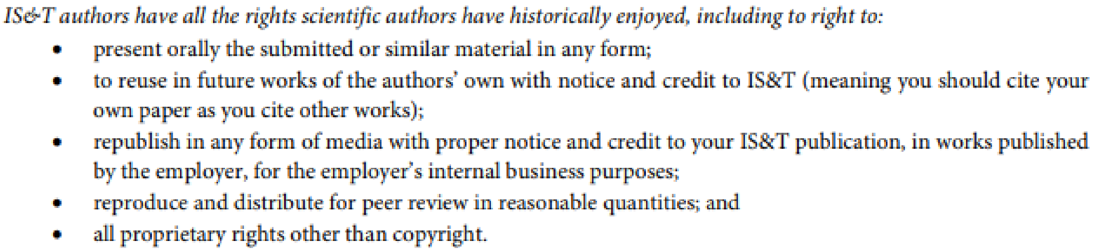
Prior versions of the Contribution published on non-commercial pre-print servers like ArXiv/CoRR and HAL can remain on these servers and/or can be updated with Author's accepted version. The final published version (in pdf or html/xml format) cannot be used for this purpose. Acknowledgment needs to be 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]".

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 IS&T

This is the permission obtained from IS&T, The Journal of Imaging Science and Technology (Chapter 4.1). The copyright page screenshot is shown in A.4.



IS&T authors have all the rights scientific authors have historically enjoyed, including to right to:

- present orally the submitted or similar material in any form;
- to reuse in future works of the authors' own with notice and credit to IS&T (meaning you should cite your own paper as you cite other works);
- republish in any form of media with proper notice and credit to your IS&T publication, in works published by the employer, for the employer's internal business purposes;
- reproduce and distribute for peer review in reasonable quantities; and
- all proprietary rights other than copyright.

Figure A.4: Copyright from IS&T