Synthesizing Micro-CT from CT of the inner ear with 3D-conditional GANs

By

Xujuan Sun

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Computer Science

May 14, 2021

Nashville, Tennessee

Approved:

Jack H. Noble, Ph.D.

Yuankai Huo, Ph.D.

ACKNOWLEDGEMENTS

My time as a graduate student has been both a personal and scientific adventure. First and foremost, I would like to thank Prof.Noble, I can’t finish the project without your help. I am deeply thankful for Jianing Wang’s help building cGAN network. I am grateful for the support of Ziteng Liu who provide the baseline for this work. Thanks for the warm and open environments at Biomedical Image Analysis for Image Guided Interventions Laboratory (BAGL).

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

I. Introduction

Cochlear implants (CIs) are neural prosthetics that are used to treat sensory-based hearing loss. Each CI comprises a microphone and processor worn behind the ear, as well as an array of between 12 and 22 electrodes, depending on the manufacturer. The implant must be surgically implanted into the cochlea, as the electrodes are designed to directly stimulate the auditory nerves and thus induce the sensation of hearing. Each electrode corresponds to a specified frequency range, and the electrode is activated when sounds in that range are detected by the microphone.

While CIs are very effective in restoring hearing, it is rare for a patient to achieve full natural fidelity. This is due to a multitude of reasons, including sub-optimal electrode positioning [1]. The position of electrodes impacts neural stimulation patterns, where electrodes positioned more distant to the modiolus create broader stimulation patterns. The region of the modiolus that is stimulated is dependent on the depth of insertion of the electrode into the cochlea. Audiologists, who in general do not have tools that permit identifying the patient specific activation regions, typically assume neural activation patterns are consistent across electrodes and patients when programming CI settings. Thus, the programming process follows a one-size-fits-all paradigm. After implementation, the audiologist attempts to optimize CIs settings including determining which electrode should be active, which stimulation level should be assigned to each electrode and so on. In recent work, we have shown that when audiologists are provided with estimations of patient-specific neural activation patterns, CI settings can be customized, and hearing outcomes can be significantly improved [2]. We call this process Image-Guided Cochlear Implant Programming (IGCIP). The approach is to use patient CT scans to localize the electrodes and neural stimulation sites. Neural stimulation patterns are then estimated based on the distance between the two. Because the neural stimulation sites are small, they cannot be directly visualized in CT images. Our group has developed image processing techniques that permit accurately localizing boundaries between intra-cochlear structures based on extra-cochlear landmarks that are visible in images [3, 4].

The initial implementation of IGCIP estimates neural stimulation patterns using the distance between electrodes and neural sites. We are developing more comprehensive physics-based electro-anatomical models to simulate patient-specific neural stimulation patterns that we believe will be more accurate than the distance-based estimate and lead to greater improvement in hearing outcomes with IGCIP [5]. Constructing the model requires estimating a high-resolution tissue class map in the region around the cochlea to delineate air, soft tissue, and bone. The model then relies on the tissue map to determine electrical resistivity throughout the volume to estimate the electric field. The resulting electric field is then used to drive neural fiber models which allows estimating patient-specific neural activation patterns. Micro-CT is an imaging modality that provides adequate resolution for constructing the tissue class map but can only be acquired for ex vivo specimens due to radiation constraints. Micro-CT thus cannot be used directly to produce a model for an in vivo target patient. The solution first proposed by our group was to use an atlas-based solution. With this approach, a set of tissue class maps, created by thresholding high resolution micro-CTs of cochlea specimens, can be registered to the target cochlea using image registration techniques [6]. Finally, the patient CT image can be segmented into a class map by using a majority vote approach with the registered micro-CT class maps.

In this work, we aim to investigate whether a deep learning based approach can provide a more accurate tissue classification map for the electro-anatomical models. Specifically, we are attempting to implement two methods. The first is using a conditional generative adversarial network (cGAN) to synthesize detailed micro-CT volumes using only clinical low resolution patient CT scan as input, followed by thresholding the resulting synthetic Micro-CT to produce the tissue classification map. The second is using cGAN do a multitask learning to generate both a detailed synthetic micro-CT and a corresponding tissue classification map simultaneously.

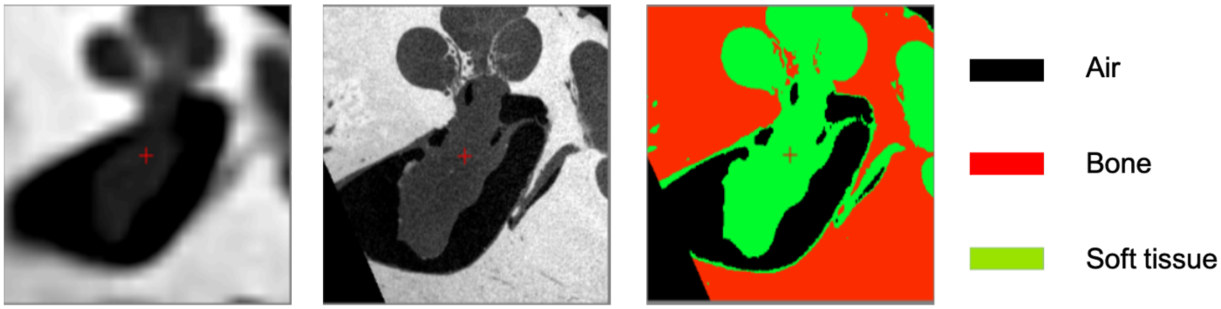


Figure 1 CT(Left), micro-CT(Middle), and tissue classification map from micro-CT(Right)

II. Related Works

Our task can be classified to an image Super Resolution (SR) problem which aims to generate high-resolution images from low-resolution inputs. Models of SR tasks can be divided into two categories: supervised and unsupervised SR models. We employed a supervised SR approach since the network will be trained with both low-resolution image and corresponding high-resolution images.

Convolutional Neural Networks (CNNs) have commonly been used for Single Image Super Resolution (SISR). Super Resolution Convolutional Neural Networks (SRCNNs) have similar architecture where the input is either the original LR image or a pre-processed image and then use hidden CNN layers to map LR image to HR image. Dong et al[7] first reported a simple SRCNN model where the input pre-processed upscaled image followed by a shallow CNNs. While as suggested by the name of Deep Learning Network, CNN models usually benefit from a deeper depth as they allow modeling mappings of high complexity[8, 9]. Kim et al[10] reported a deeper network with the use of skip connection. To efficiently train deep network architectures, the concept of residual blocks[11] and skip-connections[12][10] can be very helpful. Many research studies (e.g, [13, 14]) show that the deconvolution layers at the end of the network is the reason of checkerboard artifacts. Deconvolution layer is replaced by subpixel convolution to reduce the checkerboard artifacts. There are a lot of other promising models that have been trained with supervised learning. For example, Zhao et al [15] proposed a pyramid pooling model based on the method of spatial pyramid pooling layer [16]. Ahn et al [17] proposed a CARN model which to reduce learning complexity.

CNN models have also been successfully used in medical imaging field[18-20]. Floris et al[21] reported a deep learning-based cochlear automated segmentation system which can be used in human cochlear clinical ultra-high-resolution (UHR) CT images. There are two main algorithms in this system. The first is a segmentation algorithm that can co-learn from a computer-aided detection scheme and a U-Net-like deep learning network. The second algorithm is a measurement algorithm that can provide automatic cochlear measurements (volume, basal diameter, and cochlear duct length (CDL)).

In addition, generative adversarial nets (GANs) have been applied to many computer visions tasks and have achieved many impressive results [10]. Super Resolution Generative Adversarial Network (SRGAN) is a different model which has been shown a better perceptual result in SR tasks. Instead of focusing on minimizing the mean squared reconstructed error, SRGAN focus on how to recover fine texture details in SR tasks. Ledig et al[22] reported a SRGAN model which can generate realistic features but is lower in accuracy when compare with SRCNN model.

SRGAN model are commonly used in medical image area. Lu et al[23] published a SR classification approach which can be used to refine initial segmentation of facial nerve from low resolution of the cone-beam computed tomography(CBCT) images. The approach can be divided into two steps. An initial segmentation provided by OtoPlan[24], then a supervised learning scheme is adopted to learning the mapping between CBCT/CT images to manual created HR facial nerve segmentations.

Li et al[25] reported a cycle-consistency Generative Adversarial Networks (cycleGAN) based method which can synthesize micro-CT images from unpaired data. The Li et al reported model can be divided into two parts. The first is an unpaired image-to-image translation network used to map the CT form LR space to HR space. The second part is Bayesian inference which can be used to quantify model uncertainty. You et al [26] reported a GAN based semi-supervised deep learning approach to learn the mapping from LR CT images to HR micro-CT images. In their work, the Wasserstein distance is employed to establish the nonlinear mapping and deblurred HR outputs.

Wang et al [27] developed a method which can be used to synthesize pre-operative images from post-operative images based on GANs. In particular, conditional generative adversarial nets (cGANs) have been shown to typically perform better on image-to-image task [28, 29]. Inspired by such generative approaches, we proposed the cGAN based micro-CT synthesize method.

Current deep neural networks benefit from large scale datasets, particularly in the domain of medical image analysis [30]. However, currently there is no publicly available large dataset for evaluating cochlea micro-CTs. In order to address this problem, we used an available small dataset of micro-CT/CT pairs and employed the leave-K-out strategy for a proof-of-concept study.

2D networks are commonly used in image tasks but they can only extract two-dimensional spatial features. Though 2D networks can be applied to the volumetric data like CT image, researchers [31] have demonstrated the benefits of using 3D network to handle volumetric data. In particular, Wang et al [10] got higher image quality by a 3D network. Inspired by exploring 3D spatial context, we adopted 3D cGAN in our cochlear micro-CT synthesis task.

The rest of paper is organized as follows. In Chapter 3, we present more details about the data and the proposed method. In Chapter 4, we present the performance of our proposed method quantitatively and qualitatively. Finally, we present our conclusions in Chapter 5.

III. Materials and Methods

1. Data collection and data pre-processing

The data set includes 6 CTs and corresponding micro-CTs, where the voxel dimensions of CT are 0.3mm isotropic and the voxel dimensions of micro-CTs are 37.6um isotropic. Each CT and micro-CT pair is obtained from same ear. To create the CT and micro-CT pair, we first registered them using manual initialization and then optimized based on mutual information [32].

After registration, we created a tissue classification map with Otsu's [33] method from the micro-CT. Otsu's method employed a threshold selection approach from gray-level histogram, and we implemented the multi-threshold version. Specifically, we threshold the micro-CT to air, soft tissue, and bone. After thresholding, the tissue classification map can contain small holes. We implement a custom morphological "closing" algorithm to fill these small holes. For each class we first create a binary image and then implemented dilation algorithm followed by an image erosion algorithm. The main steps of the proposed "closing" algorithm can be seen in Figure 2.

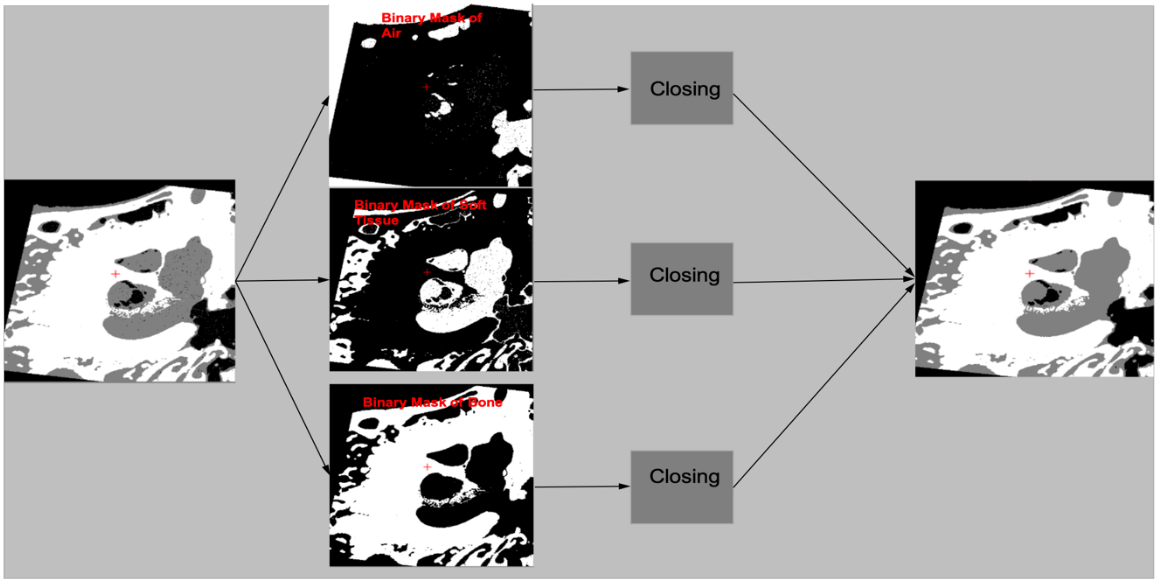


Figure 2 Main steps of "closing" algorithm. The first step is abstracting three binary images for air, soft tissue, and bone. Then, apply standard closing algorithm to close holes on binary images. Finally, combine closed binary images to get the final result. In the final combine step, if a pixel is associated with multiple class, the priority is soft tissue class over air class over bone class.

Since our data set is small, we applied the leave-k-out strategy where k is 2 in our project. After registration, we have 6 CT micro-CT pairs that are labelled with study codes S3_1, S3_2, S3_7, S3_8, S3_11, and S3_12. We divide the six pairs into 3 groups, each group has 3 pairs for training, 1 pair for validation and two pairs for testing. The detail can be seen in Table 1.

Table 1 Detailed group information for Leave-K-Out strategy

Group	Testing index	Training index	Validation index
A	S3_2, S3_7	S3_1, S3_11, S3_12	S3_8
B	S3_1, S3_8	S3_2, S3_11, S3_12	S3_7
C	S3_11, S3_12	S3_2, S3_7, S3_8	S3_1

After registered to micro-CTs, CT images are normalized to the same resolution with micro-CTs. Due to the GPU limitation, the up-sampled CTs cannot be housed into the GPU under the 3D cGANs parameters. To address the memory issue, canonically, 2D methods are used that allow networks to train slice by slice. However, 2D approaches can lose the spatial information since each slice only contain features in dimension. The second method is down-sample the CT and micro-CT pairs that networks can capture special features. However, the final goal of our project is to synthesize high-resolution CTs and producing accurate tissue classification map. The other method is to split the whole CT volume into small 3D patches. The benefit of training networks with patch strategy is two-fold. First, a truncated receptive field can reduce network complexity while still capture the local anatomical information[8, 34, 35]. Second, patch strategy can significantly increase the number of training samples[35, 36]. Deep learning benefit from big patch size since the larger the patch size is the more spatial information it contains. After testing on Nvidia Maxwell GPU, 96 * 96 * 96 is the largest patch size with a smoothing training and testing process. We have 5000 CT patches in experiments, and the training time took about two weeks. In order to reduce the training time and also to utilize spatial information across high-resolution CT and micro-CT pairs. We down-sample the CT and micro-CT with factor 2 in our multitask model and then split the down-sampled CTs and micro-CTs into small 3D patches with size 96x96x96.

We follow the standard split for training, validation and testing sets. In the training set the 3D patches are cropped with overlapping strategy where the overlapping size is 24. In the validation dataset and testing dataset the 3D patches are split without overlap.

2. Network Architecture

In this work, we employed the cGAN framework in a pixel to pixel scheme, which was proposed by Isola et al [28]. Typically, GANs [37] contain two parts, a generative network (G), which can generate fake HR images from LR ones, and a discriminative network (D), which is trained to discriminate between real and fake

HR images. D and G are trained iteratively. cGANs [38] are a special case of GANs in which both G and D are conditioned on additional information which can be used to direct the data generation process. Because our goal is to synthetize high-resolution cochlear images based on clinical cochlear images (low-resolution), the cGANs are conditioned on the clinical cochlear CT images. G thus produces a high-resolution image G (synthetic micro-CT) from the clinical cochlear CT, and the synthesized image should be “similar” enough to the target high-resolution image (micro-CT) to attempt to fool D.

Our 3D cGAN network is motivated by [10]. The framework of G can be divided into three parts. The first part contains three convolutional blocks, the second part consists of 6 ResNet blocks, followed by another three convolutional blocks. The framework of D consists of three convolutional blocks. Implementation of our network is based on Pytorch and adapted from [39]. The detailed network framework can be seen in Figure 3.

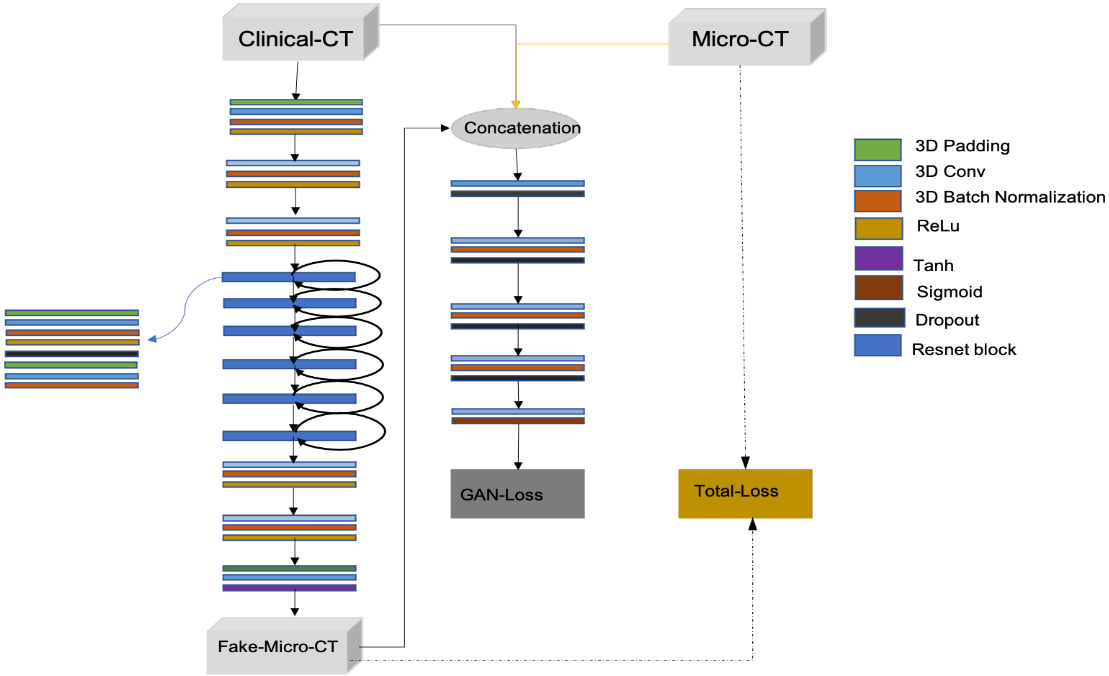


Figure 3 The illustration of 3D-cGAN single task framework

To acquire a better accurate tissue classification map, our group designed and implemented two models that are single-task model and multi-task model [40]. The single task and multitask model are implemented with same discriminator. In single-task model, the generator only generates fake micro-CTs, while in multi-task model the G generates both fake micro-CTs and fake tissue classification maps. We also applied the Dice loss on multi-task model to help the fake tissue classification maps generating process which will be discussed in loss function part. The detailed multitask network framework can be seen in Figure 4.

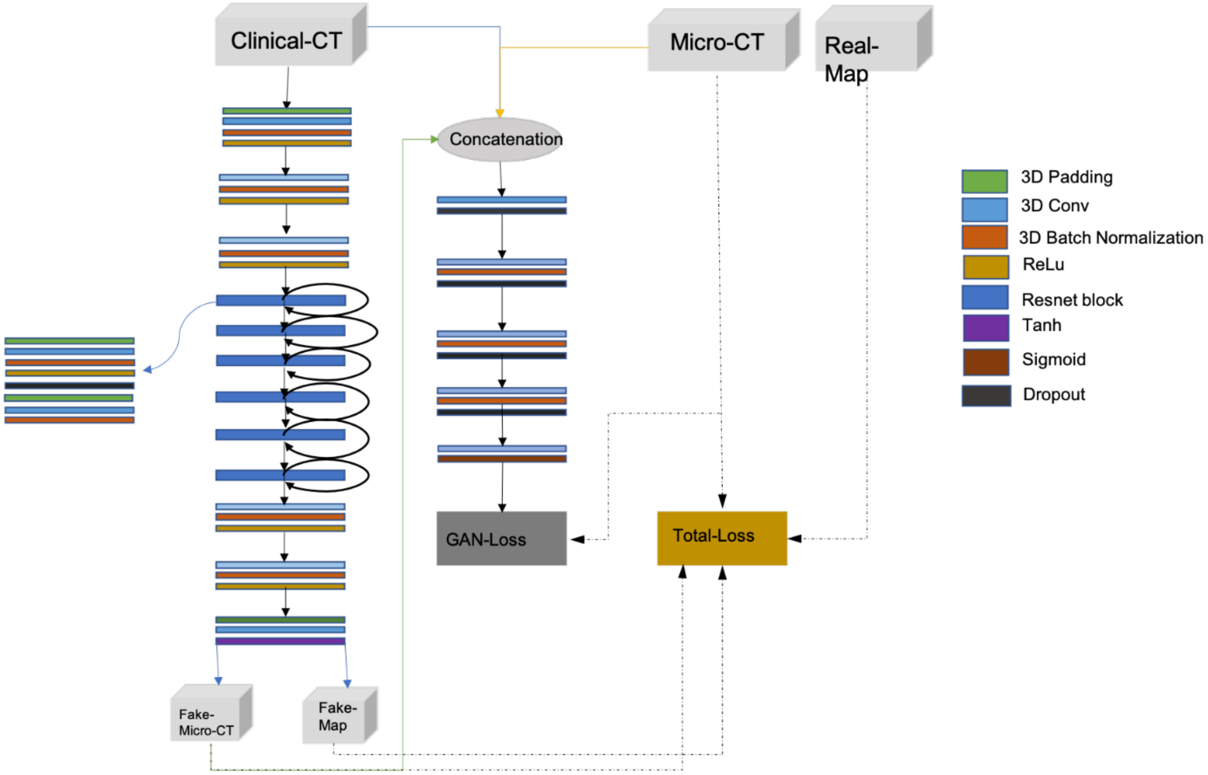


Figure 4 The illustration of 3D-cGAN multitask framework

3. Loss function and evaluation

3.1 Adversarial loss

As we discussed before, the objective of G is to fool D, and the objective of D is to assign the correct label for ground truth and the data produced by G. so the adversarial loss can be expressed as Eq.(1):

$$L_{GAN} = \underset{G}{\operatorname{argmin}}\, \underset{D}{\max}\, L_{cGAN}(G, D) \qquad (1)$$

Wherein

$$L_{cGAN}(G,D) = \mathbb{E}_{CT,micro\_CT}\left[\log\left(D(CT, micro\_CT)\right)\right] + \mathbb{E}_{CT,z}\left[\log\left(1 - D\left(CT, G(CT,z)\right)\right)\right]$$

and z is a random noise vector used by the generator G to synthesize a fake patch from the CT patch.

3.2 Weighted L1 loss

L2 distance is commonly used in networks to guild the training process, while some research studies[26, 41] point out that the output optimized by L2 norm may suffer the over-smoothing problem since L2 try to maximize the peak signal-to-noise rate[22]. L1 loss works well in previous research studies [14], so we also applied L1 loss in our task. For our ultimate goal, that is generating accurate tissue classification map, we notice that there is a heavy class imbalance problem in our input data, i.e, the number of pixels belong to soft tissue is much smaller than bone and air. Inspired by the strategy in[42] , we assign a different weight for different type of pixel. The intuition here is giving the pixel higher weight if it belongs to soft tissue. To do so, we first count the pixel number belong to soft tissue (N_{st}), bone(N_b), and air(Na), then let the weight for each class is $1/N_x$. The weighted L1 loss can be expressed as shown in Eq.(2):

$$L_{l1} = \mathbb{E}_{CT,micro\_CT}[||W \circ (micro\_CT) - G(CT))||_1] \quad (2)$$

Wherein $W = \begin{bmatrix} 1/N_a \\ 1/N_{st} \\ 1/N_b \end{bmatrix}$

3.3 Perception loss

Different from per-pixel loss between output image and ground-truth like L1 loss, Johnson et al[34] reported a perceptual loss based on high-level features extracted from pre-trained network. Johnson et al achieved a pleasing visual resulting in image super resolution task. Inspired by this work, our group also implemented perceptual loss. Instead of using a pre-trained network to extract high level feature, we use layers extracted from D directly to extract high-level feature. The perceptual loss can be expressed as Eq.(3):

$$L_{perception}(G) = \mathbb{E}_{CT,micro\_CT}[||D\_layer(micro_{CT}, micro\_CT) - D\_layer(G(CT), G(CT))||_1] \quad (3)$$

Wherein D_layer indicate the layers extracted from D.

3.4 Dice loss

Dice loss is a kind of function based on Dice coefficient. Dice loss is commonly used for segmentation tasks in deep learning to measure the overlap between two classification maps[43, 44]. In our multi-task model, G not only produces fake micro-CT but also fake classification map. We applied Dice loss for the multi-task model. The Dice loss can be expressed as shown in Eq.(4):

$$DL_u = 1 - \frac{2\Sigma_i^N p_i g_i + \epsilon}{\Sigma_i^N p_i + \Sigma_i^N g_i + \epsilon} \tag{4}$$

where G is the reference tissue classification map (ground truth) with pixel values g_i, and P is the predicted fake tissue classification map for the CT with pixel p_i, the ϵ term is used here to ensure the loss function stability by avoiding the numerical issue of dividing by 0.

To conclude, the total loss of D for single task model and multitask model are same. The total loss for D for our task is GAN loss as shown in Eq.(1).

the total loss of G for single-task model can be expressed as shown in Eq.(5):

$$L_{total\_Single} = L_{GAN} + 5 * L_{l1} + 10 * L_{perception} \tag{5}$$

The total loss of G for multitask model can be expressed as shown in Eq.(6):

$$L_{total\_multi} = L_{GAN} + 5 * L_{l1} + 5 * L_{perception} + 10 * DL_u \tag{6}$$

4. Evaluation

The proposed method is compared to the method proposed by our previous study. Since the ultimate goal is generating a high accuracy tissue classification map. The result of our model is compared with previous method in Dice similarity coefficients (DSC). DSC can be expressed as shown in Eq.(7). The method of getting tissue classification map from single task model is thresholding the fake micro-CTs generated by G with Otsu's method followed by the "closing" algorithm. While in the multitask model, fake tissue classification maps can be generated directly from G.

$$\mathrm{DSC} = \frac{2|\mathrm{A} \cap \mathrm{M}|}{|\mathrm{A}|+|\mathrm{M}|} = \frac{2|\mathrm{TP}|}{2|\mathrm{TP}|+|\mathrm{FP}|+|\mathrm{FN}|} \tag{7}$$

Wherein TP is true positive, FP is false positive, FN is false a negative.

Our models also generate fake micro-CTs. To visually evaluate fake micro-CT generated from single task model compare to multitask model can be difficult if the differences are small. Inspired by the work of [45] , our group extend the method used in [45] to qualitatively compare the visual result of single task model and multitask model. The perceptual error in this paper can be defined as shown in Eq.(8).

$$P_{err} = \frac{sum(^{|micro\_CT-G(CT)|}/_{micro\_CT})}{micro\_CT.size} \tag{8}$$

Wherein $micr\_CT.size$ is the volume size of micro-CT.

IV. Experiment and result

1. Experiment details

The input of G is a 1-channel 3D CT patch with size 96*96*96. The output of G is corresponding fake CT or fake CT and fake tissue classification map depending on the networking framework. The input of D is the concatenation of micro-CT patch and it's corresponding fake CT. As suggested by Isola et al[28], drop out strategy is applied in our task to introduce randomness when G is generating fake micro-CTs.

Our models are trained alternatively between G and D, i.e., we train one step on G and then one step on D. Adam optimizer with momentum 0.5 is applied in our single task and multitask model. The batch size is 1, and the epoch number is 200. In the first 100 epochs, the learning rate is fixed as 0.0002, in the second 100 epochs, the learning rate is reduced to 0 linearly. The model is saved in every 5 steps, and the fake micro-CTs generated in the training process are uploaded to website for better monitoring the training process.

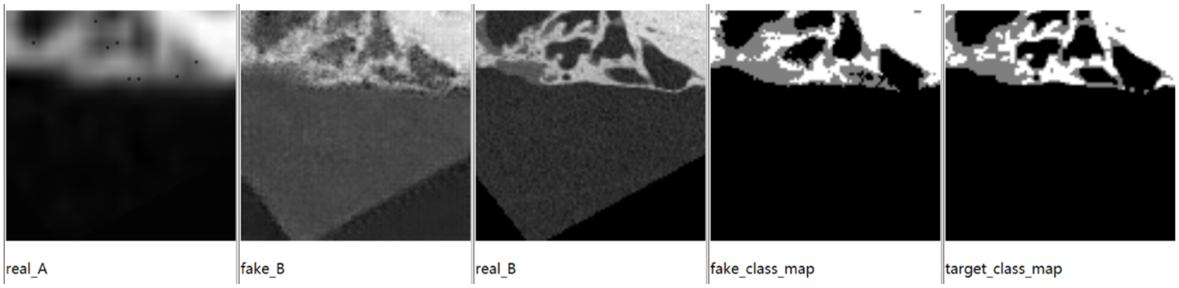


Figure 5 Shortcut of the website used for monitoring the training process. real_A is the input CT, fake_B is the fake micro-CT generated by G, real_B is micro-CT, fake_class_map is the fake tissue classification map generated by G, and target_class_map is tissue classification map(In single task model, only real_A, fake_B, and real_B will show up in the website)

2. Validation

We applied leave-k-out strategy in our task as discussed above. We have 3 groups. In each group we have a training dataset, validation dataset, and testing dataset. After the training process, we tried to find the epoch with lowest loss value by applying our validation dataset every ten epochs from 0 to 200. We did the validation process for all the 3 groups. The best network epoch number for each group are shown in table 2, and the Group_A' s loss curve for a single task model is shown as Figure 6.

Table 2 Network epoch number for each group

	Group_A	Group_B	Group_C
Single task model	135	115	135
Multitask model	95	95	135

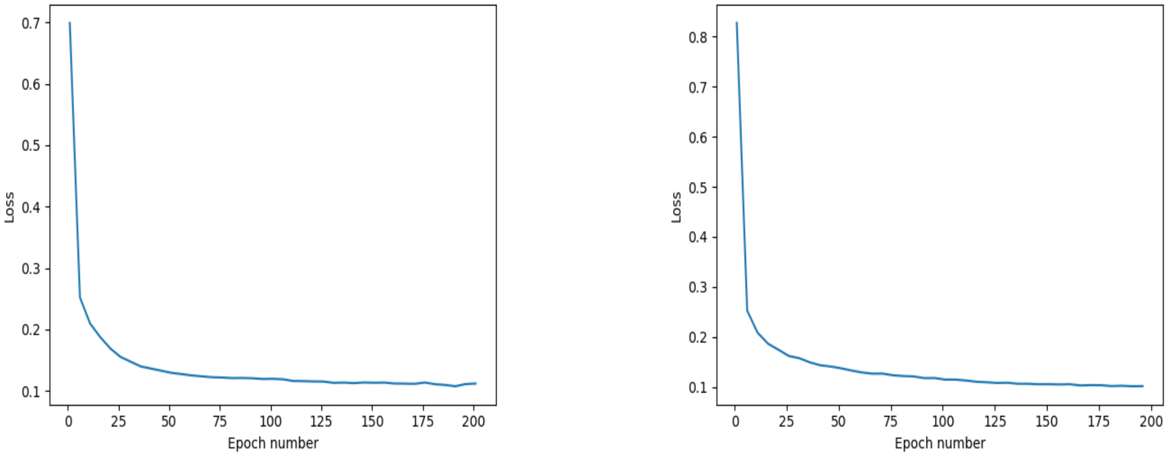


Figure 6 Group_A' s validation loss curve in single task model(left) and multitask model(right)

For the single task model, the losses used to plot the validation curve are weighted L1 loss and perception loss since we are more interested in the visual result in single task model. For the multitask model, the losses used to plot the validation curve are weighted L1 loss and Dice loss since we are more interested in the class map in multitask model.

3. Results

After the validation process, we tested each group with the network chosen by the validation process. In order to compare the results between different network models and the previous atlas-based method, we compute the Dice score. Detailed result can be seen in Figure 6:

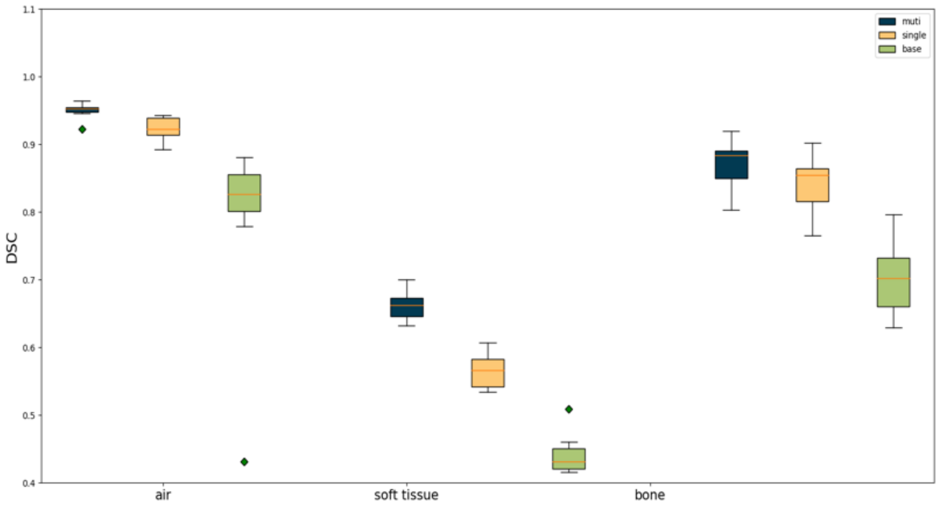
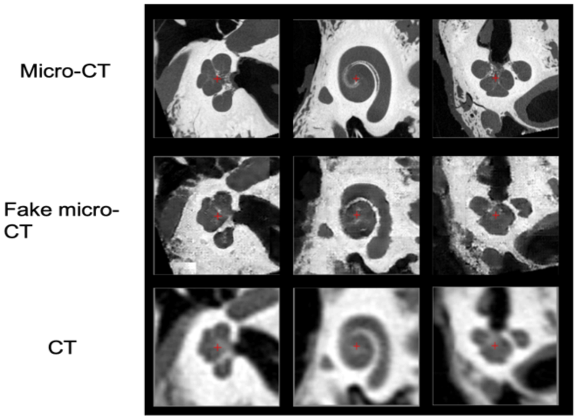


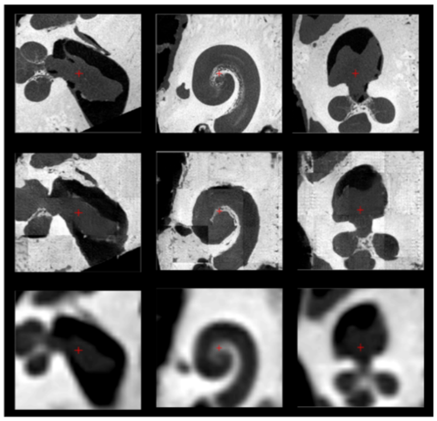
Figure 7 Boxplot of Dice Similarity Coefficient (DSC)

Figure 7 shows the boxplots of Dice score for the 6 specimens. Figure 6 shows that both single task method and multitask model obtained better results than our previous method. For class air, the median of our previous method is 0.83, the medians of the single task model and multitask model are 0.92 and 0.95. For class soft tissue, the median of our previous method is 0.43, the medians of the single task model and multitask model are 0.57 and 0.66. For class bone, the median of our previous method is 0.68, the medians of the single task model and multitask model are 0.85 and 0.88. Both single task model and multitask model achieve better results against our previous method on Dice score, and the Dice scores of multitask model are higher than single task model.

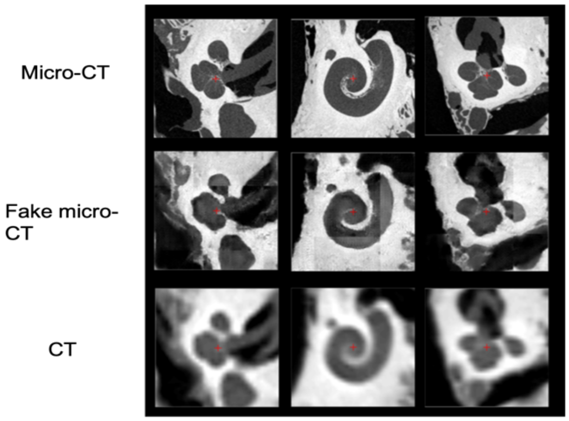
Clinically relevant structure cannot be evaluated by Dice scores. We also compare the visual result between single task model and multitask model. The detailed visual result can be seen in Figure 8.



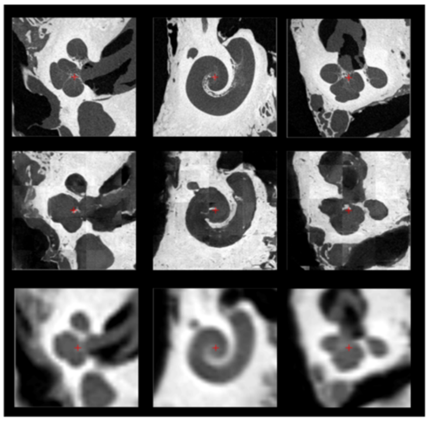
Multitask result for S3_1



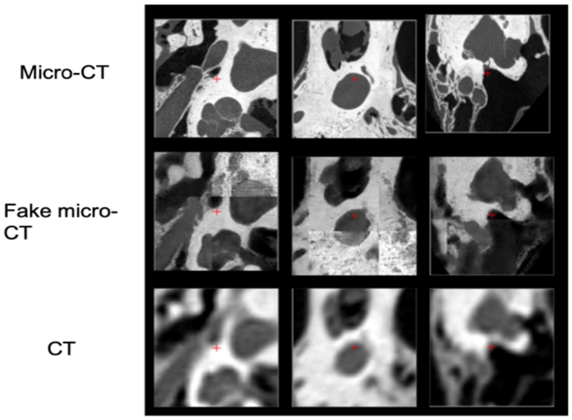
Single task result for S3_1



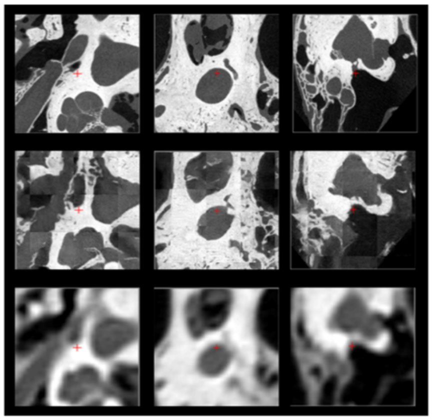
Multitask result for S3_2



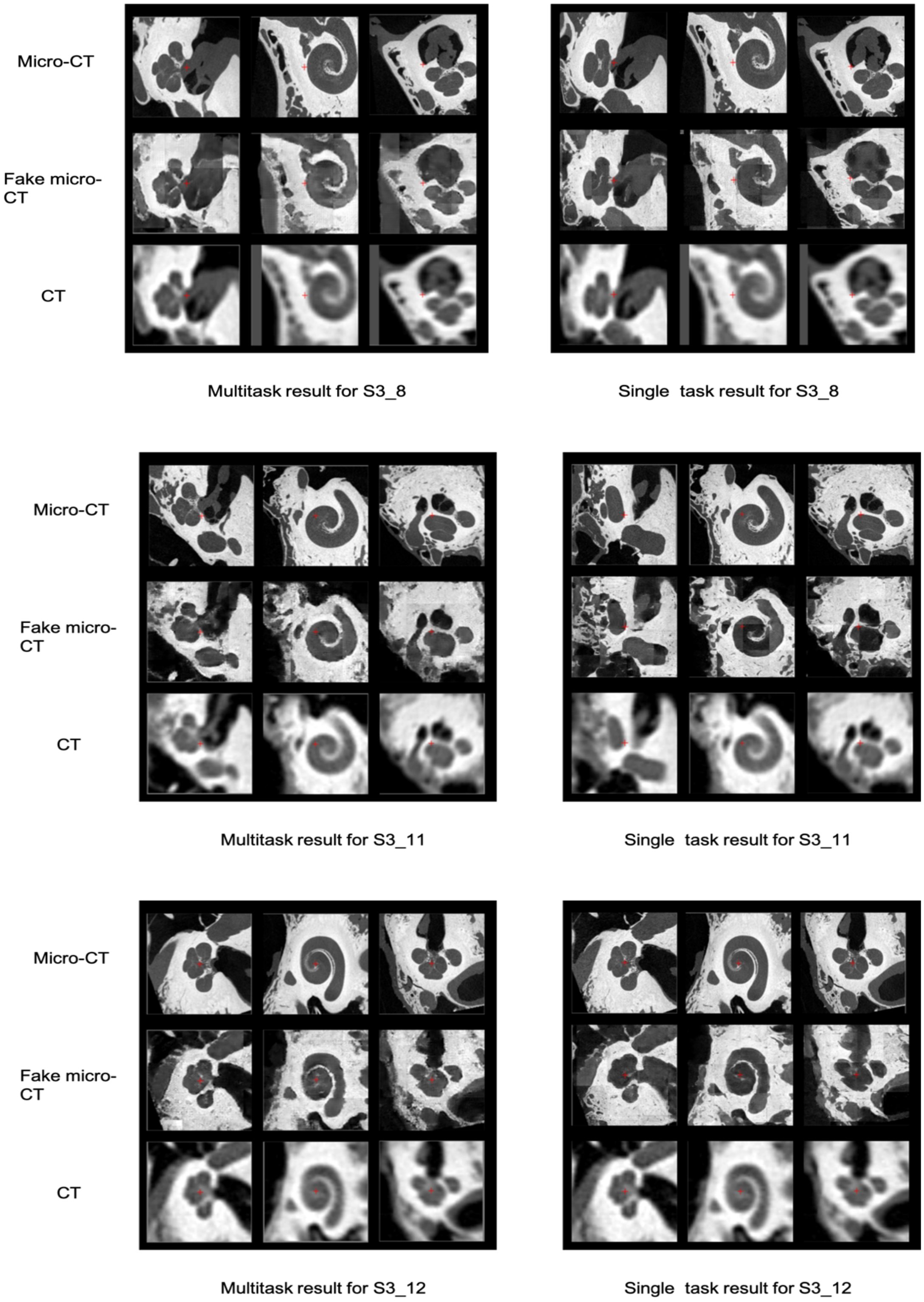
Single task result for S3_2



Multitask result for S3_7



Single task result for S3_7



Multitask result for S3_8

Single task result for S3_8

Multitask result for S3_11

Single task result for S3_11

Multitask result for S3_12

Single task result for S3_12

Figure 8. Visual results for both single task model and multitask model

Form Figure 8, we can see that single task model can generate a better visual result since the fake micro-CT generated by single task model has a more consistent boundary and image contrast among different patches. In order to better compare the visual results between single task model and multitask model as discussed above, our group calculate the perceptual error for both single task model and multitask model, detail can be seen in Table 4.

Table 3 Perceptual error for both single task model and multitask model

Specimen index	Multitask model	Single task model
S3_1	2.18	2.17
S3_2	5.1	1.8
S3_7	6.5	8.6
S3_8	1.8	1.8
S3_11	3.1	2.8
S3_12	2.4	2.1

As shown in table 3, the perceptual error of single task model is smaller than the error in multitask model expect for S3_7. In general, the single task model has a better visual result than model task model.

V. Discussion

In this study, two 3D-cGAN based systems were developed to synthesize the micro-CTs from CTs. In the single task system, micro-CTs are acquired from network G, and the corresponding tissue classification maps can be achieved with Otsu's thresholding method. In the multitask system, both micro-CTs and tissue classification maps can be acquired from network G. Both single task model and multitask model were compared with our previous Atlas-based method. The visual result only compared between the single task model and multitask model since our previous Atlas-based model can only generated tissue classification map.

The GAN based architecture enable the system the ability of generating detailed cochlear structures which can't be obtained from clinical CTs. Some researches[25] are reported to use cycle-GAN to generate micro-CTs. However, at the time of writing and to our best knowledge, this is the first time cGANs were used to synthesize both tissue classification maps and micro-CTs from clinical CTs. GAN-based models are usually hard to train since GANs contain two networks. This problem is more challenge for our work due to a small dataset. Leave-K-out and data augmentation[46] are two commonly used methods for the small dataset problem in deep learning. Thus, we implement a Leave-K-out strategy in this paper. We also augment our dataset by splitting CT volumes into small, overlapping patches.

Our study has shown that it is possible to use cGAN based method to synthesize micro-CTs and create more accurate tissue classification maps as presented in the results. However, we believe there is still room for improvement. In the single task model, topological loss function can be implemented to add more shape constrains in the training process and make the output more coherent in a global sense[47]. In the multitask model, Dice loss was implemented to guide the generation of tissue classification. Hausdorff Distance(HD)[48, 49] is another criteria which is commonly used in deep learning segmentation task and considered as a more informative indicator. It is worth to try HD loss[50] in future work. As for the 3D patch strategy, overlapping was only implemented in the training process in our current work. It also worth to try the method of making the patches overlapped with each other in the validation and testing process. With the overlapped patches in validation and testing process, whole synthesized volume can be reconstructed through averaging the overlapped patches, and this could address the patch boundary consistency artifacts.

REFERENCES

[1] L. K. Holden *et al.*, "Factors affecting open-set word recognition in adults with cochlear implants," *Ear and hearing,* vol. 34, no. 3, p. 342, 2013.

[2] J. H. Noble, R. F. Labadie, R. H. Gifford, and B. M. Dawant, "Image-guidance enables new methods for customizing cochlear implant stimulation strategies," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 21, no. 5, pp. 820-829, 2013.

[3] J. H. Noble, R. F. Labadie, O. Majdani, and B. M. Dawant, "Automatic segmentation of intracochlear anatomy in conventional CT," *IEEE Transactions on Biomedical Engineering,* vol. 58, no. 9, pp. 2625-2632, 2011.

[4] J. H. Noble, R. H. Gifford, R. F. Labadie, and B. M. Dawant, "Statistical shape model segmentation and frequency mapping of cochlear implant stimulation targets in CT," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2012: Springer, pp. 421-428.

[5] A. Cakir, R. T. Dwyer, and J. H. Noble, "Evaluation of a high-resolution patient-specific model of the electrically stimulated cochlea," *Journal of Medical Imaging,* vol. 4, no. 2, p. 025003, 2017.

[6] J. V. Hajnal and D. L. Hill, *Medical image registration*. CRC press, 2001.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence,* vol. 38, no. 2, pp. 295-307, 2015.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[9] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[10] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646-1654.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*, 2016: Springer, pp. 630-645.

[13] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill,* vol. 1, no. 10, p. e3, 2016.

[14] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874-1883.

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881-2890.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 37, no. 9, pp. 1904-1916, 2015.

[17] N. Ahn, B. Kang, and K.-A. Sohn, "Image super-resolution via progressive cascading residual network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 791-799.

[18] G. Wang, M. Kalra, and C. G. Orton, "Machine learning will transform radiology significantly within the next 5 years," *Medical physics,* vol. 44, no. 6, pp. 2041-2044, 2017.

[19] G. Wang, "A perspective on deep imaging," *Ieee Access,* vol. 4, pp. 8914-8924, 2016.

[20] G. Wang, J. C. Ye, K. Mueller, and J. A. Fessler, "Image reconstruction is a new frontier of machine learning," *IEEE transactions on medical imaging,* vol. 37, no. 6, pp. 1289-1296, 2018.

[21] F. Heutink *et al.*, "Multi-Scale deep learning framework for cochlea localization, segmentation and analysis on clinical ultra-high-resolution CT images," *Computer methods and programs in biomedicine,* vol. 191, p. 105387, 2020.

[22] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681-4690.

[23] P. Lu *et al.*, "Highly accurate facial nerve segmentation refinement from CBCT/CT imaging using a super-resolution classification approach," *IEEE transactions on biomedical engineering,* vol. 65, no. 1, pp. 178-188, 2017.

[24] N. Gerber, B. Bell, K. Gavaghan, C. Weisstanner, M. Caversaccio, and S. Weber, "Surgical planning tool for robotically assisted hearing aid implantation," *International journal of computer assisted radiology and surgery,* vol. 9, no. 1, pp. 11-20, 2014.

[25] H. Li *et al.*, "Micro-CT Synthesis and Inner Ear Super Resolution via Bayesian Generative Adversarial Networks," *arXiv preprint arXiv:2010.14105,* 2020.

[26] C. You *et al.*, "Structurally-sensitive multi-scale deep neural network for low-dose CT denoising," *IEEE Access,* vol. 6, pp. 41839-41855, 2018.

[27] J. Wang, J. H. Noble, and B. M. Dawant, "Metal artifact reduction for the segmentation of the intra cochlear anatomy in CT images of the ear with 3D-conditional GANs," *Medical image analysis,* vol. 58, p. 101553, 2019.

[28] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125-1134.

[29] J. Lin, Y. Xia, T. Qin, Z. Chen, and T.-Y. Liu, "Conditional image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5524-5532.

[30] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems,* vol. 24, no. 2, pp. 8-12, 2009.

[31] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis,* vol. 42, pp. 60-88, 2017.

[32] D. Zhang, R. Banalagay, J. Wang, Y. Zhao, J. H. Noble, and B. M. Dawant, "Two-level training of a 3D U-Net for accurate segmentation of the intra-cochlear anatomy in head CTs with limited ground truth training data," in *Medical Imaging 2019: Image Processing*, 2019, vol. 10949: International Society for Optics and Photonics, p. 1094907.

[33] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics,* vol. 9, no. 1, pp. 62-66, 1979.

[34] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, 2016: Springer, pp. 694-711.

[35] J. Hamwood, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, "Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers," *Biomedical optics express,* vol. 9, no. 7, pp. 3049-3066, 2018.

[36] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794-2802.

[37] I. J. Goodfellow *et al.*, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661,* 2014.

[38] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784,* 2014.

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223-2232.

[40] R. Caruana, "Multitask learning," *Machine learning,* vol. 28, no. 1, pp. 41-75, 1997.

[41] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging,* vol. 3, no. 1, pp. 47-57, 2016.

[42] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. of the Int'l Conf. on Artificial Intelligence*, 2000, vol. 56: Citeseer.

[43] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, 2016: IEEE, pp. 565-571.

[44] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*: Springer, 2017, pp. 240-248.

[45] P. Welander, S. Karlsson, and A. Eklund, "Generative adversarial networks for image-to-image translation on multi-contrast MR images-A comparison of CycleGAN and UNIT," *arXiv preprint arXiv:1806.07777,* 2018.

[46] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics,* vol. 10, no. 1, pp. 1-50, 2001.

[47] J. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. King, "A topological loss function for deep-learning based image segmentation using persistent homology," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2020.

[48] W. R. Crum, O. Camara, and D. L. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE transactions on medical imaging,* vol. 25, no. 11, pp. 1451-1461, 2006.

[49] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC medical imaging,* vol. 15, no. 1, pp. 1-28, 2015.

[50] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on medical imaging,* vol. 39, no. 2, pp. 499-513, 2019.