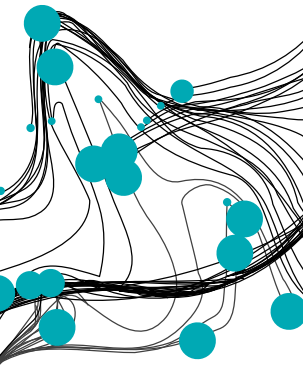
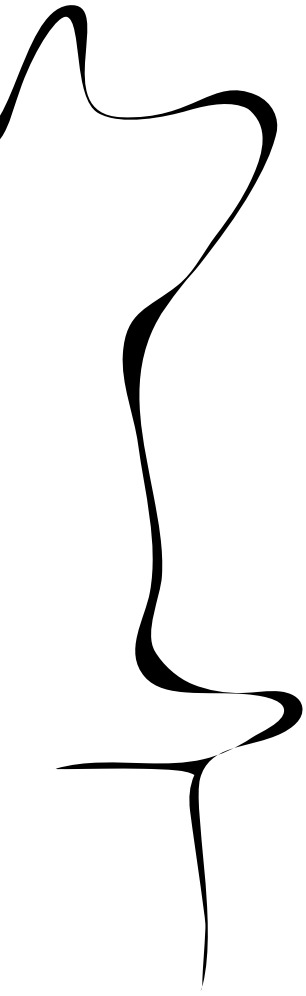


RAM ROBOTICS AND MECHATRONICS



GENERATION OF LUNG CT IMAGES USING SEMANTIC LAYOUTS

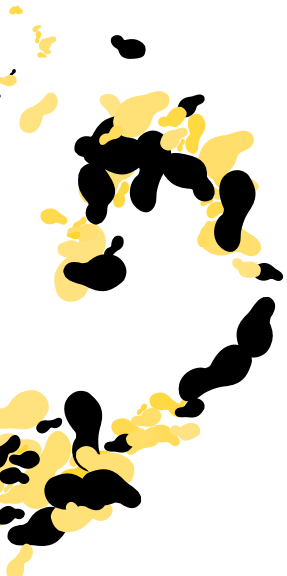
S. (Sheng-Chih) Wu

MSC ASSIGNMENT

Committee:
dr. ir. F. van der Heijden
E.I.S. Hofmeijer, MSc
dr. ir. L.J. Spreeuwers

August 2021

057RaM2021
Robotics and Mechatronics
EEMathCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands



UNIVERSITY OF TWENTE. | TECHMED CENTRE

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Summary

Lung cancer is one of the most severe cancers, and early diagnosis is important for life-saving interventions. However, the diagnosis of the lung cancer, which is based on repetitively monitoring axial lung Computed Tomography images and manually segment the nodules, is tedious and the accuracy are heavily depends on the experience of radiologists. Many researchers and developers have been working on computer-aided diagnosis (CAD) system to improve the productivity of the radiologists. As the development proceed, the radiologists need to learn how to use these tools efficiently to support their daily work, and a computer assisted learning (CAL) system is required for training and assessing the radiological professionals in diagnosis with CAD system. Ideally, the CAL system could render diverse and realistic patient cases that the radiologists can practice on it, but it requires innumerable amount of data to fulfill that requirements.

Compared to other well known datasets, medical datasets are relatively small and it is not easy to acquire sufficient medical data due to the privacy issues. To tackle this issue, Generative Adversarial Networks are widely explored and continuing developing in medical field. Recently, researches have focused on conditional synthesis models to generate images in uncommon conditions and improve the robustness of the model. Moreover, some researches aim to enable user-interactive manipulation, by which users are able to guide the generation of synthesis images.

In order to explore the usability of state-of-the-art method, we utilised Spatially Adaptive Denormalization (SPADE) to synthesize the lung CT synthesis images. Several metrics for fidelity evaluation were used to score the quality of synthetic medical images. We validated these metrics by aligning the results with visual examination, and we found that Frechet inception distance (FID) could better reflect the quality. Moreover, we trained pix2pixHD, another semantic image synthesis method to compare with SPADE. As the result, SPADE did not surpass pix2pixHD in quality measurements, but it outperformed the other in the degree of multi-modal synthesis, meaning that SPADE was able to render varied outcomes with slightly lower quality. We also tested the manipulability of SPADE by editing nodule masks. By relocation, expand and shrink, we found that the texture of nodules could be synthesized differently. We further applied the manipulation on the semantic labels to generated synthetic images for data augmentation, and the experiments showed that the augmented data improved the performance of segmentation network in dice coefficient and sensitivity. We also compared the SPADE-based data augmentation method with the common method which the data was augmented with flipping ground truth, and the results showed that the synthetic method did not surpass the traditional method.

The results show that SPADE can generate a realistic lung CT images, but there are some limitations to be tackled. For fidelity measurement, the visual examination was done without domain experts. To further evaluate the fidelity, a perceptual study conducted by experienced radiologists is recommended. The validity of FID in lung CT images can be also verified by the study. For inducing diversity, the variety of synthetic nodules are limited compared to the ground truth, since SPADE will generalize the attributes of synthesis results. To overcome this challenge, it is recommended to classify the nodule into several classes based on their attributes such as subtlety. By dividing nodule class into several semantic classes, the attributes information can be preserved during SPADE learning process, and the SPADE can synthesize each class specifically. Overall, theses results suggest that SPADE is an effective method to medical image generation due to its quality and manipulability, and it can be further develop as either CAL system for training junior professionals or data augmentation for the medical imaging.

Contents

1 Introduction

1.1 Motivation

Lung cancer is one of the most severe cancers. It is the leading cause of cancer deaths worldwide [1], and early diagnosis is crucial for successful treatment and life-saving interventions [2]. The diagnosis of lung cancer is based on quantification of pulmonary nodules using axial lung Computed Tomography (CT) image. The non-invasive imaging technique that constructs the internal structure of human body by 2D scans. Since pulmonary nodules can be associated with several diseases, continuous monitoring and accurate segmentation of pulmonary nodules are required for malignancy estimation and forecast [3] [4]. However, these are tedious works which introduces the inter-observer variabilities [5].

To overcome these challenges during manual segmentation, researchers have been working on implementing computer-aided diagnosis (CAD) systems to alleviate the work load and enhance the productivity of radiologists. Several techniques have been proposed to support radiologists, including image processing based techniques [6]. In order to implement these techniques successfully, a computer assisted learning (CAL) system is required for training and assessment of radiological residents and technical physicians in diagnosis with advanced CAD systems. More importantly, it needs to be able to adjust cases to the needs of the trainee.

Compared to other well known datasets, medical datasets are relatively small. It is not easy to acquire sufficient medical data due to the privacy issues. That is, patient consent may be required to use the diagnostic images. The data scarcity will limit the capability of artificial intelligence, and slow down the research and industrial progress. To tackle this issue, Generative Adversarial Networks [7], which have the power to generate photorealistic images are widely explored and continually developed in the medical field. This is useful to generate images in uncommon conditions, improving the robustness of the model [8]. Moreover, some researches aim to enable user-interactive manipulation, by which users are able to guide the generation of synthesis images [9].

To enable user-interactive manipulation, the research community has explored conditional image synthesis, which includes image translation from text, labels or images. In order to provide a user-friendly editing tool, a semantic image is used as conditional input. However, concatenating semantic label to input to the generator does not work, since the conventional normalization layers tend to wash away this information [10]. Semantic aware approaches are necessary [11] for developing such an image generation tool. Following the method proposed by Park et al., by redesigning the normalization layer to be spatially-adaptive and semantic-aware, the semantic information can be preserved and used as style guidance to supervise the generation process of synthesis images. Combing with PatchGAN discriminators [12] or segmentation networks, it has been shown that semantic image synthesis is possible for scenes such as urban streets and landscapes. In this project, we would make use of these ideas for semantic image synthesis and explore the usability of it to generate realistic lung CT images with an adjustable degree of complexity and solve the problems.

1.2 Related Work

1.2.1 Medical Image Synthesis

GAN (Generative Adversarial Networks) have been used in the medical research community for style translation [13], super-resolution [14], segmentation [15] and data augmentation [16]. Chuquicusma et al.[17] have proposed to use a DCGAN to synthesize realistic lung nodule and evaluate them by visual Turing tests, the first time in the literature utilising the GANs to generate lung nodules. Qin et al. [18] utilise the GANs to synthesize volumetric lung nodules which

will be used as data augmentation for training the segmentation network. The results show that by using GAN-based augmentation, the performance of the segmentation network improves. Yang et al. [19] proposed a method based on WGAN with a perceptual objective that can reduce the noise level and reconstruct the CT images, which have been taken under low dose condition.

There is not much research in medical GAN application utilising semantic label map. Proposed by Tony et al. [20], the semantic label map was used to enhance the brain tumour generation under elastic deformation. In order to generate detail content, they proposed a multi-task generator to reconstruct the tumour and other parts separately. The proposed generator is used to synthesize rare cases where the brain is deformed largely. The method is proposed as a data augmentation technique improved from traditional elastic deformation, while the capability of semantic manipulation remains unclear. Besides, inputting semantic label map directly to the generator, according Park et al., is not an optimal way for semantic image synthesis. Although there are many state-of-the-art approaches proposed, it was not until recently that the concept of semantic image synthesis became popular. The community has not explored the possibility of implementing semantic image synthesis in the medical field yet. This motivates us to explore the usability through this project.

1.2.2 Conditional Image Synthesis

Generative adversarial network has become popular recently due to its capability in image synthesis. Applying GAN with conditional setting is referred as Conditional Image Synthesis [21][22], which GAN synthesizes based on given conditions. Many researches have explored different conditions to constraint the GAN synthesis, from labels [23] [24], text [25] to images. Specifically, the image-conditioned models refer to those aimed to learn mapping from source domain to target domain. See Figure 1.1. Denoting the source domain as A and the target domain as B, the goal of these models is to learn the mapping $G_{A\to B}$, translating the input image x_A to x_{AB}, where $x_{AB} \in B$. In the other words, the models convert the extrinsic style representation of x_A to B while preserving the intrinsic source content. Compared to L_1 loss, which usually leads to blurry images, applying trainable adversarial loss, which can learn and adapt the difference between the source and target domain, automatically has become popular for image-to-image translation. For example, the pix2pix framework has been used to transfer Google maps to satellite view [12]. Researchers have explored the usability in different applications, such as future frame prediction [26], style transfer [27] [28], supperesolution [29] and photo manipulation [9].

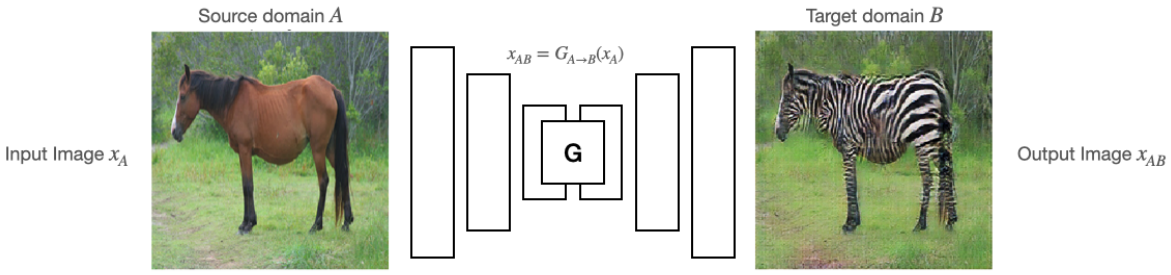


Figure 1.1: Image-to-Image translation

1.2.3 Semantic Image Synthesis

Recently, some works have focused on using a semantic label map as condition for interactive image manipulation, later called semantic image synthesis. Directly inputting semantic label

map to conventional image-to-image translation models does not give the promising results, due to the loss of semantic information during normalization. To tackle this challenge, GauGAN [10] has been proposed to preserve the semantic conditions by Spatially-Adaptive Denormalization (SPADE). Instead of passing through the normalisation, SPADE utilises the semantic mask as condition of modulation parameters. By such modification, the semantic label maps can be used to guide the synthesis process, enabling semantic image synthesis. Following the success of SPADE, many researches have focused on improving the discriminator structure. The OASIS [11] model is proposed to replace the PatchGAN [12] discriminator with segmentation network, given a better alignment between synthesis images and semantic label map. The SESAME [30] model proposed that the semantic label maps and RGB images should be input to the discriminator separately for better preservation. Besides the conditional batch normalisation method, CC-FPSE [31] proposed a conditional convolution generator in which the convolution kernels are conditioned on semantic label maps through a weight prediction network. So far, semantic image synthesis has not been used in synthesizing lung CT images, and it remains unknown how suitable these models are in the medical imaging domain.

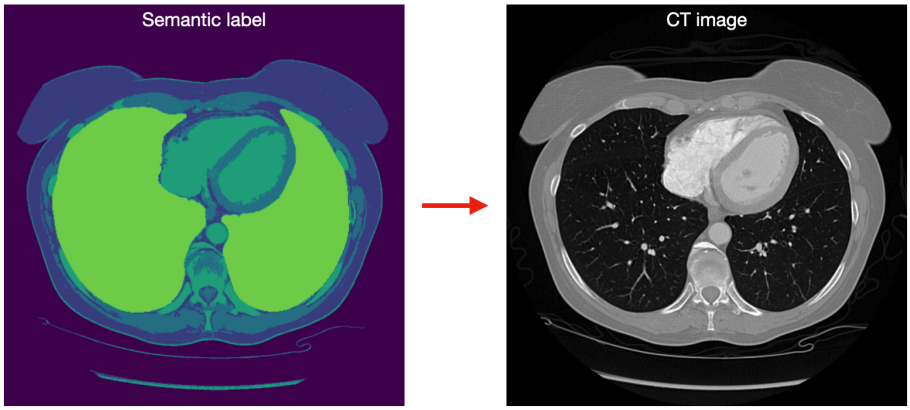


Figure 1.2: Semantic Image Synthesis

1.3 Objective

The goal of this project is to implement a semantic image synthesis model to generate lung CT images. The main research question will be:

To what extent can the semantic synthesis GANs be applied to lung CT images?

We further develop several sub questions, aiming to give a comprehensive study of the selected semantic image synthesis model for lung CT images. The sub-questions that will help to answer the main research questions are:

- How perceptual realistic are the generated images?
- To what degree can we influence the outcome by user input?
- To what extent can the generated images used for data augmentation?

We evaluate the fidelity of generated images following standard GANs protocol, following the evaluation metrics of semantic image synthesis papers. Due to the capability of semantic image synthesis, we also study the influence of manipulation on synthesized images by user input.

There are several challenges in order to answer these questions. First, there is no research conducted using a semantic image synthesis model in the medical field yet. Although there are

many state-of-the-art models proposed recently, most of them follow the precedents, using the same datasets like COCO-Stuff [32] and Cityscapes [33]. There is no baseline or a proper reference for usability in medical images. Therefore, we define the project as an exploration research, aiming to investigate the topic which is not studied yet. Second, most of the lung CT image synthesis researches focus on the region of interest (ROI) only, so there is no annotations for other tissues in the chest. To overcome this obstacle without support from the experts, it is necessary to design an image processing algorithm to generate the semantic label from input data. Therefore, the design of such algorithm will also be a key part of this exploration project. We aim to set up a process and generate the semantic label of full CT images with adjustable parameters, such that it can be applied and extended for future works. Third, since there is no metric that has been proven effective in the medical image domain, we utilise common metrics of semantic image synthesis papers. The evaluation results will be correlated with visual examination, which will be done by human without domain expertise.

1.4 Report Outline

The remained of this report is structured as follows. Chapter 2 will provide background information of the selected semantic image synthesis model. The structure and the theory behind the model will be explained in order to better understand the state-of-the-art techniques. Chapter 3 will cover the method, mainly the design of the image-processing algorithm for semantic label generation, the materials, the training details and the experiment set-up. The results of the experiments will be shown in Chapter 4, and the discussion of the results will be in Chapter 5. Finally, based on the results and discussion, the conclusion of the exploratory research will be drawn in Chapter 6, with the proposed future work.

2 Background

In this chapter, we will first review the literature of GANs in medical imaging, from the general concept behind GANs to its variants. There are many GANs implementations in medical imaging such as segmentation [34], and we specifically focus on image synthesis [13].

Since SPADE is referred as conditional image synthesis, the difference between unconditional and conditional image synthesis will be explained, aiming to give a comprehensive overview of the emerging fields. After that, we narrow down to SPADE, describing the detail of the architecture, and the design choice it made to enable semantic image synthesis. Since SPADE is an improved model from several precedents, we summarise those literature thoroughly to provide a concise review.

2.1 Generative Adversarial Networks

GAN [7] is a generative model trained via adversarial process, in which a generative model G captures the data distribution and generates samples, and a discriminative model D estimates the probability that a sample comes from training data rather than G.

The GANs process can be described as follows. Consider a training set X, a generator G with parameters θ_G, a discriminator D with parameters θ_D, and a random noise z. G aims to map $\hat{x} = G(z;\theta_G)$ for which $\hat{x} \in \hat{X}$, and the primary goal is to optimise the mapping such that $p_\theta(\hat{x}|z) \sim X$, meaning that the generated samples $\hat{X}$ resembles the distribution of the training set X. The mapping is validated by discriminator D, which classifies between fake and real samples, yielding $D(x;\theta_D) = 1$ for real data and $D(x;\theta_D) = 0$ for generated data.

The adversarial process is achieved by the interaction between both networks, as G attempts to synthesize realistic samples that can fool D, while D continuously learns to differentiate between real and generated one. The feedback is achieved by gradient information propagated back from D to G, and G adapts its parameters to produce a better output. The adversarial process can be described mathematically as two-player minimax game with value function $V(G,D)$:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[\log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[1-\log(D(G(z)))] \tag{2.1}$$

Through optimisation, D is trained to maximise the probability of correct classification between real and fake data, while G is trained to generate realistic data by minimising $\log(1 - D(G(z)))$. The optimisation is achievable theoretically by minimising the Jensen-Shannon (JS) divergence between the distribution of X and $\hat{X}$, but it has been proven to be hard to train due to training instability and mode collapse.

2.2 GANs in Medical Image Synthesis

Originally, GAN is proposed as an unsupervised generative framework. It is able to map from random noise to the realistic synthetic images. In cGAN[21], it has become a supervised generative framework for which the synthesis process is influenced by conditional information instead of random noise alone. The original GAN, known as vanilla GAN, is referred as unconditional GAN [8][35] in contrast to the conditional GAN.

There are several applications of high quality image synthesis in the medical field. First, the synthesis images can be used as augmented data for training the deep learning network. With conditional settings, it is possible to alleviate the data scarcity and class imbalanced problems. Second, the image-to-image translation models can be implemented as domain transfer for different imaging modalities. Several researches have been motivated for this purpose and shown promising results. Third, if the generated samples are realistic enough, it is possible to

use the generative adversarial models as CAL system, the educational tool to train junior professionals. Fourth, it can improve diagnostic accuracy by mining highly discriminative imaging features.

2.2.1 Unconditional Synthesis

Unconditional synthesis indicates image generation from noise without auxiliary information. It has become popular in medical field since it could tackle data scarcity problems, which are considered the major challenges in medical imaging. Researches have shown that unconditional GANs could synthesize realistic medical images that the human experts could hardly distinguish from the real one. [36] utilises DCGAN [37] to learn mapping the brain MRI to the implicit manifold, and synthesizes high resolution brain MRI with only small amounts of training data. The synthesis results are visual compelling which two imaging experts could not be distinguish reliably from the real brain MRI. This indicates deep convolutional networks have potential to synthesis photorealistic images with small amounts of medical data. [17] has also utilised DCGAN to generate lung nodule. It is the first literature using deep learning network to synthesize class-specific lung nodules. Two types of nodules, benign and malignant, are trained separately, and a series of Visual Turing tests are executed with two radiologists to quantitatively evaluate the synthesis results. The results show that the radiologists could not reliably tell the difference between the generated and real nodules for both types. The analysis also shows the intra-observer variation in diagnostic lung cancers between participants. The discriminative imaging features learned by the DCGAN could help improving the diagnostic.

For full lung CT image synthesis, [38] has utilised progress growing GAN [39] to generate realistic body CT samples in high resolution. Ranging from thorax to abdomen, the slices are synthesis in an unsupervised manner and validated by a series of visual Turing test. With ten radiologists identifying synthetic samples as fake, the results show no significant difference in the specificity between the visual Turing test and random guessing. They even split the participants based on the experience level, and found that the accuracy between each group is not significantly different, but it has a limitation in synthesizing the slices of thoracoabdominal junction and the detail of anatomical texture.

2.2.2 Conditional Synthesis

Conditional GANs in medical application is used mostly for cross-modality synthesis due to clinical needs. A common situation occurs where two imaging modalities render accessorial information so that two obtainments are required for diagnostic procedure. For example, CT is used for gel-dosimetry in radiation oncology, so it has to be obtained additionally to diagnostic planning MR [40]. However, conducting CT scan could put patients at risk of cell damage and cancer due to radiation exposure [35], and this motivates the implementation of image-to-image translation techniques for cross-modality synthesis. Researches have explored synthesis CT to MR using CycleGAN [41], lung MR to CT with additional tumour-aware loss function [42], and lung PET to CT by modified cGAN [43].

2.2.2.1 Semantic Image Synthesis

In this research, the main focus will be image synthesis conditioning with semantic label, which is known as semantic image synthesis. The benefit of conditioning on semantic label is easy-editing, allowing user to control image generation by adding, removing, or manipulating the semantic label [9]. Conditioning on manipulated semantic label can render a powerful tool for medical image synthesis application.

Several researches in the medical imaging have been done by semantic image synthesis. [18] utilises cGAN to generate 2D lung nodule conditioning with semantic label. The synthesis nodules are then concatenated into the train set of proposed 3D segmentation network. The au-

thors have shown that cGAN conditioning with semantic label could be an effective method for data augmentation. [20] proposes a coarse-to-fine generative models for data augmentation of brain tumour. Compare to the brain, tumour is relatively small and the detail tends to be blurred during synthesis process. To enhance the tumour synthesis quality, additional boundary-aware generator is added. A multi-task generator is proposed to simultaneously infer the location and the boundary of the complete tumour. Compare to traditional data augmentation method such as flipping and rotation, the proposed method could render a better results for elastic deformation, providing the rarely-seen samples for the segmentation network to learn. To be noted, the authors do not mention mode collapse issues, which is commonly seen in semantic image synthesis according to the literature [10], and the effect of user manipulation for such model remains unknown.

For semantic lung CT image synthesis, [44] has proposed a conditional GAN to generate lung samples for COVID-19 diagnostic. They utilise the labels from COVID-19 dataset [45], which contains 3 classes: lung region, ground-glass opacity and consolidation. Without generating additional labels for non-lung region, the 3-classes segmentation maps are used as conditions for image generation, resulting in synthesising lung only images. The results are merged with non-lung area from the ground truth and validated as full CT images. The authors have shown that the proposed method outperforms the state-of-the-art models either with complemented images or lung only images. While it is arguable to formulate CT images by combining synthesis and ground truth partially, the boundary of non-lung region will not change correspondently with the manipulated labels. Considering the purpose of our project, all classes should be editable by users. Therefore, we did not follow the proposed method.

2.3 SPADE

In the following section, the detail of SPADE model will be explained. SPADE model, so called GauGAN model, is a method to process the semantic information for conditional generative adversarial network, aiming to resolve the mode collapse problems which happen frequently when conditioning on semantic labels.

2.3.1 Spatially Adaptive Denormalization

Spatially adaptively denormalization (SPADE) is the conditional normalization methods proposed by Park et al [10]. Compared to unconditional normalization layers, this methods rely on external data to tune the modulation parameters, given the controllable parameters that the synthesis process can be conditioned on. The general process operates as follows. The layer activations are normalized to unit deviation and zero mean. The normalized activations are then denormalized by modulating the activations based on the affine transformation, which is learned from the external data. This approach has been applied to style transfer tasks, for which the affine parameters are used to control the global style of generated images. However, instead of globally uniform affine transformation, the modulation parameters are redesigned to be channel-wise and element-wise, making it a spatially-varying affine transformation which is suitable for image synthesis from semantic maps.

The process of SPADE is similar to Batch Normalization. The activation is normalized and then modulated with the scale and bias. The semantic maps are first projected to an embedding space and then convolved, producing the scale and bias learned from the input. Those parameters and are then multiplied and added to the normalized activation element-wise. The element-wise and channel-wise manners ensure the style of synthesis objects been generated coherently, and the preserved semantic information increases the controllability of the synthesis process. The design of SPADE residual blocks largely followed ResBlock proposed by Miyato et al. [46], for which applying residual learning for super-resolution tasks. In case that the input and output have different numbers of channels, the skip connection is also learned.

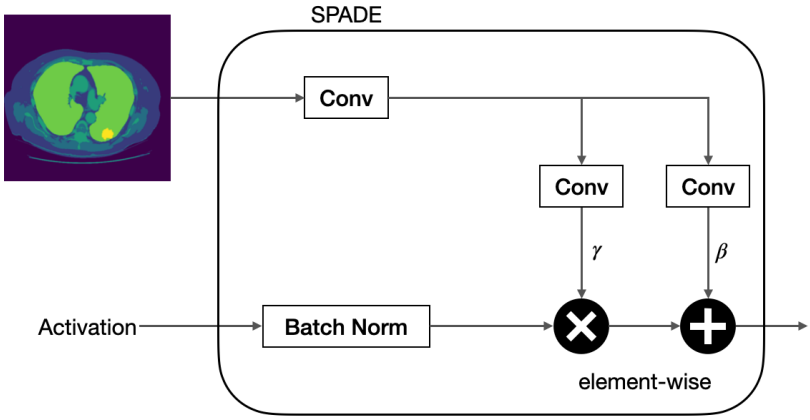


Figure 2.1: Working principle of SPADE.

2.3.2 Generator

The design of SPADE generator and discriminator largely adopt pix2pixHD [25] model with several modification. While the generator of pix2pixHD model consists of U-Net [26] like encoder-decoder network, the SPADE generator is consisted of a series of the SPADE residual blocks with upsampling layers. Compare to pix2pixHD, the encoder part of the generator is discarded while several SPADE residual blocks plugged in the decoder structure, resulting in a more lightweight network. The segmentation labels are not directly fed into the first layer or intermediate layers. Instead, the input labels are fed into the SPADE residual blocks to encode information of label layout. The generator itself only takes random vector as input.

It has been shown that the uniform appearance of semantic label will be lost by applying normalization, generating identical outputs without detailed texture. Many well-known image-to-image translation models fail to preserve the semantic information as they concatenated the labels map directly to the intermediate layers without further adjustment to ensure the information being adopted correctly. Unlike conventional ways, the SPADE generator is modulated by segmentation semantic maps from the user in different scales, adopting to the spatial resolution of residual blocks. Since the semantic maps are used in different scales through the generator pipeline, it enables coarse-to-fine generation. By such architecture, the labels maps are encoded in the spatially varying modulation parameters without normalization, so the information of semantic labels can be preserved. This turns out allowing users to control the synthesis process of the generator, enabling semantic editing of GAN generated images.

2.3.3 Discriminator

PatchGAN [12] models the image as a Markov random field, assuming independence between pixels distanced more than a patch diameter. Instead of predicting classes probabilities, PatchGAN renders classification results of divided patches from the original images. The original design of PatchGAN is to restrict the discriminator to only model high-frequency structures in local image patches, as low frequencies can be handled by L1 loss easily. PatchGAN tries to classify whether each $N \times N$ patch in an image is real or fake. The results of each patch are averaged to produce final output. Compared to conventional neural networks, applying PatchGAN encourages the generator to produce synthesis images with better local structures and contents, resulting higher quality results.

In order to generate high-resolution image, the discriminator needs to have a larger receptive field to differentiate high-resolution real and synthesized images. This would require a deeper

network or larger convolutional kernels, which are both computational expensive. Following pix2pixHD, SPADE utilises the structure which is called multi-scale discriminators. Three discriminators are created with identical network structure but different operational scales. The discriminators are trained to differentiate real and synthesized images at three scales from coarse-to-fine. While the coarse one has the largest receptive field, focusing on guiding the generator to produce global consistent images, the finest one encourage the generator to generate finer details.

2.3.4 Multi-Modal Synthesis

Since the random vector provide stochastic input to the generator, by using a random vector as the input to the generator, it enables multi-modal synthesis. By attaching an image encoder before the generator, it can be further extended to a style guidance network and form Variational Autoencoder. Namely, the VAE encodes the real image into a latent representation. By calculating mean and variance from the encoded information, the original style can be passed to the generator as random vector. The overall SPADE structure is shown in Figure 2.2.

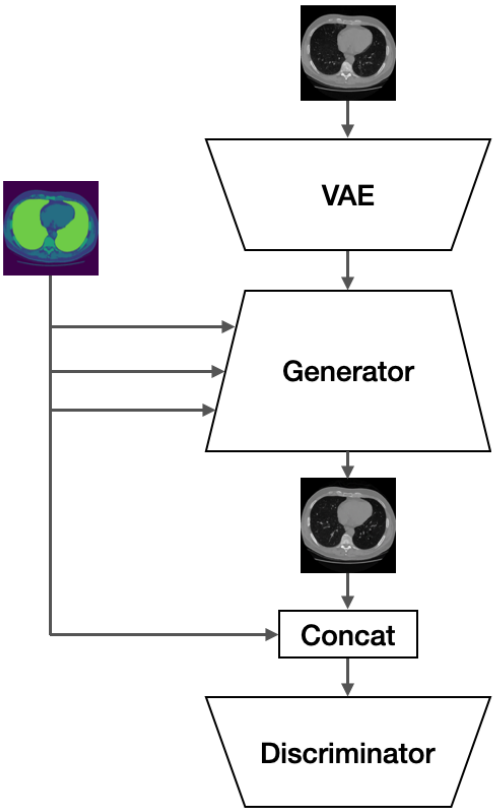


Figure 2.2: Training the SPADE with the semantic label map. The ground truth image is used as style guidance fed into the VAE.

2.3.5 Learning Objective

The Hinge loss is applied as adversarial function. As L1-term loss function tackles the correctness of low frequency content construction, PatchGAN focus on high frequency structure.

$$L_D = -E_{(x,y)\sim p_{data}}[\min(0, -1 + D(x,y))] - E_{z\sim p_z}[\min(0, -1 - D(G(z),y))] \tag{2.2}$$

$$L_G = -E_{z\sim p_z, y\sim p_{data}} D(G(z),y)) \tag{2.3}$$

Feature matching loss is also applied as improved adversarial loss. The features are extracted from multiple layers of the discriminator. By matching these intermediate representation from real and synthesis images, the discriminator learns how to distinguish them. Penalising the feature matching loss in L1 distance stabilises the training for multiple scale synthesis.

$$L_{FM}(G, D_k) = E_{s,x} \sum_{i=1}^{T} \frac{1}{N_i} [\| D_k^{(i)}(s, x) - D_k^{(i)}(s, G(s)) \|_1] \tag{2.4}$$

For k represents the amount of scale, T the feature maps from the discriminator and the N_i the normalized factor for each feature map.

The perceptual loss is also applied jointly, since it has shown improving the performance slightly. It is calculated based on the feature extracted from the intermediate layers of VGG-19 network pretrained on Imagenet. VGG-19 network has been applied in medical image synthesis to calculate loss.

$$L_{VGG}(G) = E_{s,x} \sum_{i=1}^{N} \frac{1}{M_i} [\| F^{(i)}(x) - F^{(i)}(G(s)) \|_1] \tag{2.5}$$

$F^{(i)}$ denotes the i-th layer with M_i elements of the VGG network.

The importance of each learning objects are weighted by the parameters that user inputs.

The encoder can be trained via parameterisation trick. In the GauGAN, the KL divergence loss is used for the style encoder.

$$L_{KLD} = D_{KL}(q(z|x) \| p(z)) \tag{2.6}$$

The prior distribution $p(z)$ is a standard Gaussian distribution and the variational distribution q is fully determined by a mean and a variance vector.

3 Method

In this chapter, the methodology of the experiments will be explained. We started with the generation of semantic labels. Several processes and design choices of creating 2D lung CT images dataset will be described in section 3.1. The details of the experiment designs will be illustrated in section 3.2. The training details of semantic image synthesis are in section 3.3.

3.1 Semantic label generation

The goal of semantic label generation is to extract the data from the database, adjust data for nodule analysis and generate corresponding semantic labels. The details of the utilised database will be illustrated in section 3.1.1, and the pre-processing steps will be described in section 3.1.2. From section 3.1.3 to 3.1.5, the generation of semantic label for non-lung, lung and nodule are explained respectively. Finally, the post-process step will be described in section 3.1.6.

3.1.1 Database

The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) establishes a public database for medical imaging research community [47]. It was initialised by the National Cancer Institute (NCI), aiming to help the development of CAD technologies for lung nodule classification, detection and segmentation. The LIDC-IDRI Database is collected and organised by seven academic centers and eight medical imaging companies, resulting in 1018 cases from 1010 patients. For each case, it contains images from clinical thoracic CT scan and a corresponding XML file that includes the nodule information.

The nodule is defined through a two-phase image annotation process performed by four experienced thoracic radiologists. In the initial blinded phase, each radiologist reviewed CT scans independently, and classified the lesions into three categories: nodule $\geq$ 3 mm, nodule $<$ 3 mm, and non-nodule $\geq$ 3 mm. Nodules, of which with diameter larger than 3 mm, are annotated additionally with the contours and marked attributes such as subtlety, sphericity, and malignancy. In the subsequent unblinded phase, the radiologists revised their own labels along with the anonymised marks of the other three participants to give the final determination. The purpose of such process was to identify all lung nodules in each CT scan, considering the intra-observer variabilities.

3.1.2 Pre-Processing

3.1.2.1 Rescale and Intercept

The clinical thoracic CT scans are stored in standard Digital Imaging and Communication in Medicine (DICOM) format [48], which is a widely adopted data type for communication and management of medical imaging information. Since the LIDC-IDRI Dataset was collected from different institutes, the equipment from different vendors were used to acquire CT scans. In each scan, pixel values of CT images are measured in Hounsfield units (HU), which is a quantitative scale for describing radiodensity.

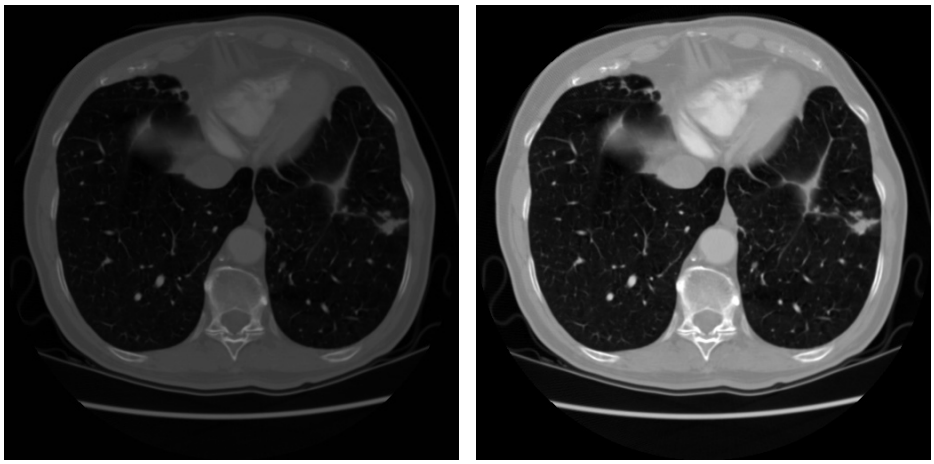
However, CT data acquired from different equipment is stored differently, and the range of intensity values may differ between the representation stored on disk versus in memory, depending on the system specification. For example, the HU values could be negative, and are usually stored as an unsigned integer with specific slope and intercept depending on vendor's preference. Therefore, it is common for CT data to have negative intercepts, and the linear scaling is applied to store the values in a memory efficient way and at the same time avoid quantisation error.

The first step after retrieving pixel values from DICOM data is to transform from memory representation to disk representation. In other words, to normalise CT data from different vendors. The linear transformation between the intensity values (IV) and Hounsfield Units (HU):

$$HU = IV \times s + i \tag{3.1}$$

where s indicates the rescale slope and i denotes the intercept. Two DICOM tags can be used to perform such transformation, rescale intercept (0028|1052), and rescale slope (0028|1053) [49].

HU values represents the relative density of tissues and organs in CT images. In order to synthesis lung CT images for radiologists to diagnose lung cancer, we followed the literature [50] to clip the pixel values to [-1024,800], which is the meaningful pixel value for nodule judgement. The results are shown in Figure 3.1.



(a) Before HU clipped (b) After HU clipped

Figure 3.1: Lung CT images before and after HU clipped.

3.1.2.2 Slice selection

In order to synthesize CT images from upper to lower chest, we included the non-nodule slices in the train set.

Compared to the number of slices that contain nodules, the amount of non-nodule slices is larger. We selected 20% of the slices from each scan, including nodule and non-nodule slices, to avoid further imbalance. The non-nodule slices are selected randomly and uniformly from each scans. We excluded CT scans whose slice thickness is greater than 2.5 mm in consideration of image quality [18]. For semantic image synthesis, we utilised the nodules which are annotated by at least two radiologists. The purpose is to acquire nodules as much as possible with at least 50% of confidence. Hence, there 49992 slices in total.

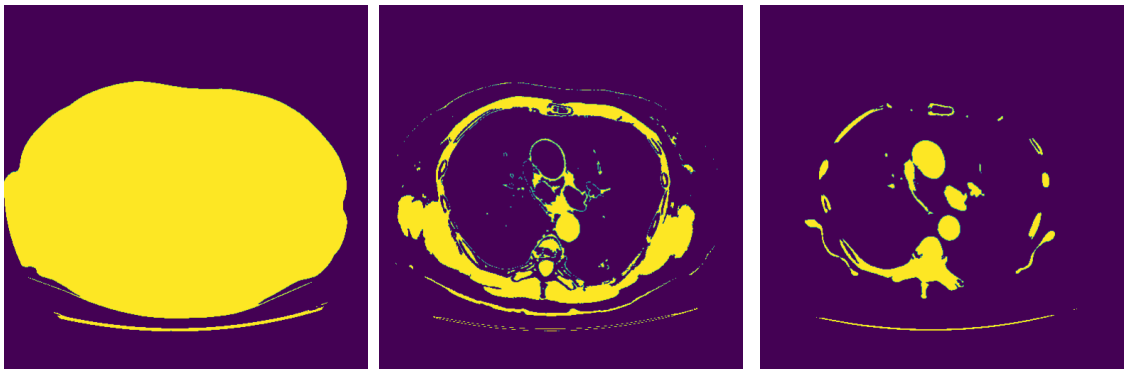
3.1.3 Non-lung segmentation

Since we did not have the explicit annotation of those non-lung tissues, we classified them into three categories based on their HU values. We found it difficult to distinguish the organs and tissues by HU values alone. For example, it is difficult to distinguish the bone from the heart since they could have the same pixel values, and so does the lung region with the soft tissue. Therefore, we segmented the non-lung substance into three classes: body, soft tissues, and high dense tissues.

The first class, the body, is to segment the thorax from the background, and it can be done robustly by segmenting HU [-400,800]. The second class, the soft tissues, refers to those substances with HU higher than the skin and fat, but lower than bone and heart. It can be organs,

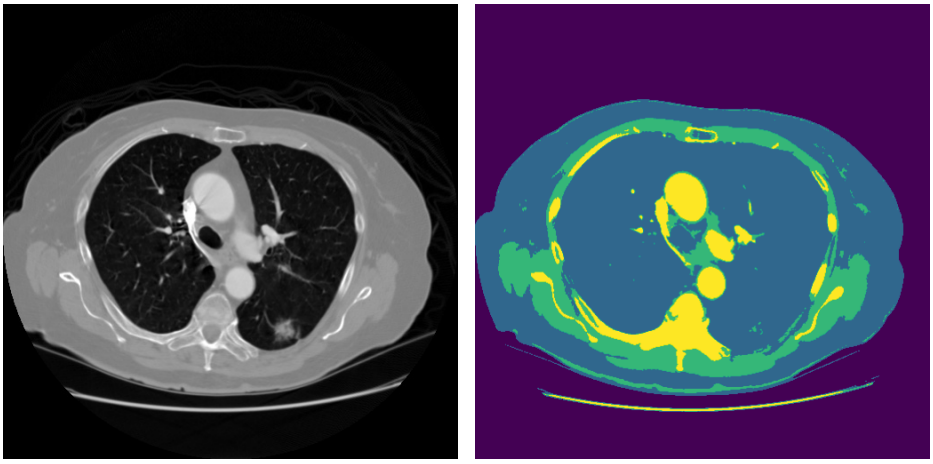
muscle and other tissues. We segmented those substance by HU [0,200]. The third class, the high dense tissue, is segmented based on the HU ranging from 200 to 800, shown as Figure 3.3. The rib cage and some muscular organs are included in this class. The pixel which belongs to both class 1 and class 2 or 3 will be overwrote by the latter, see Figure 3.

Even though we did not distinguish the organs and tissues explicitly in semantic label, we believe the semantic image synthesis model could learn synthesizing the substances based on their spatial information and surrounding.



(a) Class1: body (b) Class2: soft tissues (c) Class3: high dense tissues

Figure 3.2: Semantic labels of body, soft tissues and high dense tissues



(a) Ground Truth (b) Non-lung semantic label

Figure 3.3: Ground truth image and non-lung semantic labels

3.1.4 Lung segmentation

For the lung region, simple segmentation based on HU does not work due to several tissues with different HU vales are included. Instead, we utilised classic imaging processing techniques to segment the lung. In this research, we used K-Means classification method and marching square algorithm to generate the masks of lung.

Compared to non-lung regions, pixel values of lung area are usually lower. We used K-Means method to classify the non-lung and lung pixel with several morphology operations. The morphology operations included erosion and dilation, which were implemented to erode the bright substance in lung and expand boundary to cover the lung area as completed as possible. The background was excluded by setting up a maximum area threshold.

While K-Means method could robustly locate the lung region, it could not render a precise boundary between lung and non-lung regions. It would result in an incomplete lung mask with some substance not included, and an over expanded boundary that overlaps the other regions. Therefore, we utilised marching square algorithm to find the explicit contour of the lung. This method would depict the outlines based on the constant valued pixels surrounding an object, providing accurate contours. The depicted contours would then be transferred into masks and be compared with the masks generated by K-Means method iteratively, to filter out the exact lung contours, as shown in Figure3.4. We used cosine similarity to compare the K-Means masks with the contours efficiently. The selected contours will be transformed into the lung masks and be used as semantic label for lung. See Figure3.5(b).

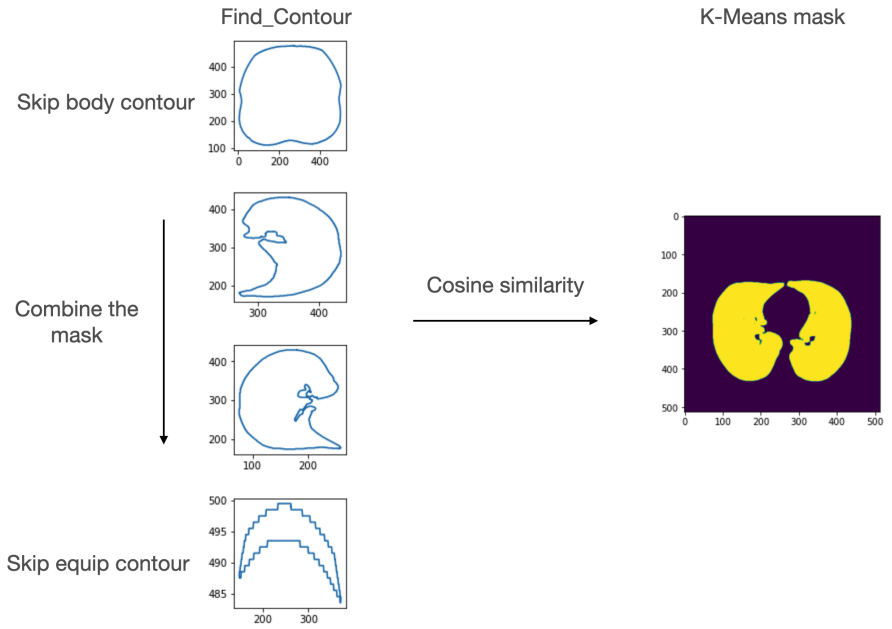
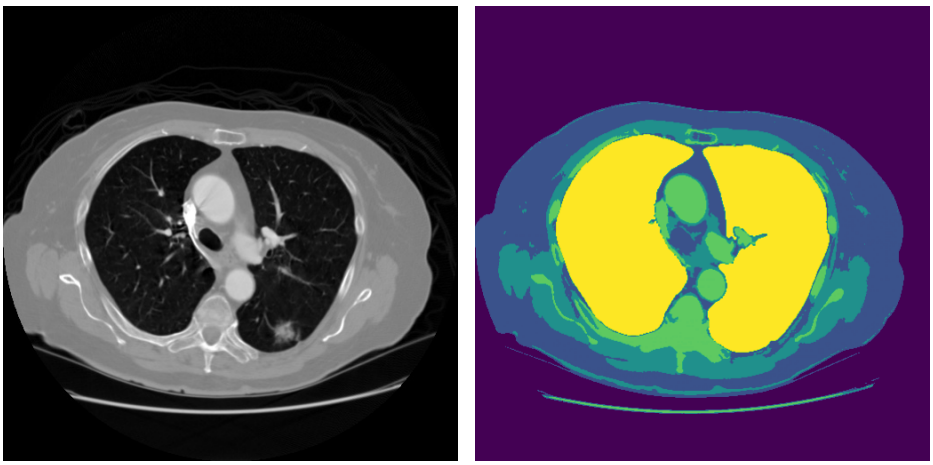


Figure 3.4: The lung mask depicted by K-Means method is used to select the lung contour by the marching squire to render a precise boundary.



(a) Ground Truth (b) Lung semantic label

Figure 3.5: The lung CT images and the semantic label embedded with lung region.

3.1.5 Nodule depiction

The semantic label of nodules were depicted by radiologists' annotation. For each experiment, the nodules annotated by the selected amount of radiologists will be used and depicted based on a 50% consensus criterion. The criterion implies that the given pixels have been included in the boundary by different radiologists, and the result is depicted at a 50% consensus level, as shown in Figure3.6. Since the annotation from different radiologist might differ, generating the nodule label at a 50% consensus level would be fair design choice. The nodule label will be used to generate the semantic label. See Figure3.7(b).

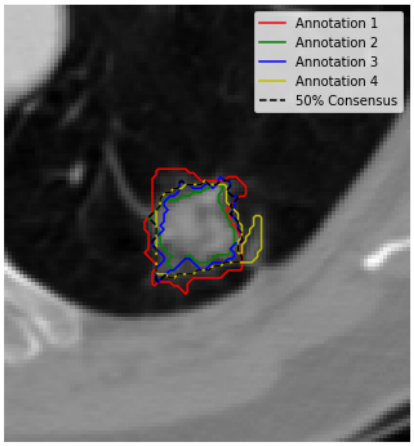
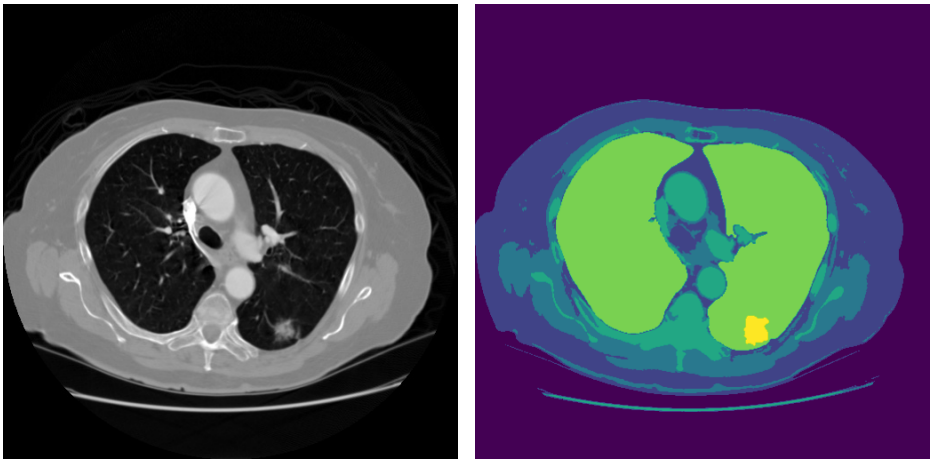


Figure 3.6: Nodule depiction with 50% consensus.



(a) Ground Truth (b) Lung semantic label

Figure 3.7: Lung CT images and the full semantic labels.

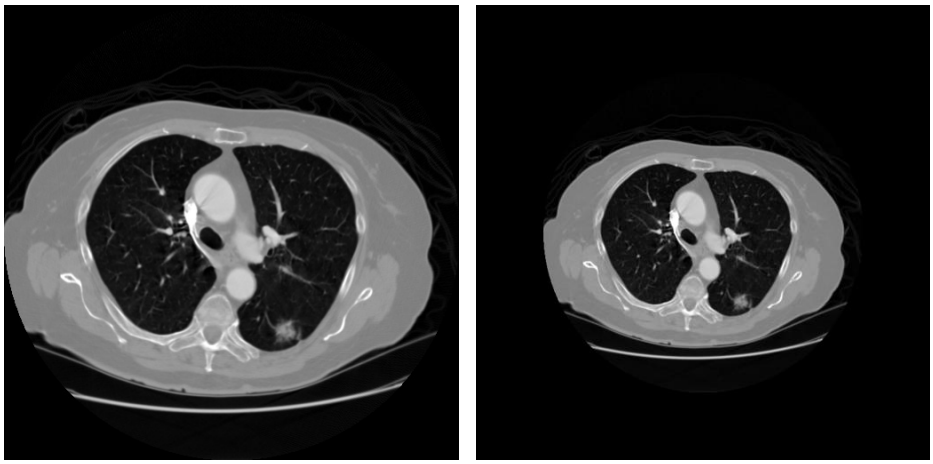
3.1.6 Resampling

The pixel spacing attribute (0028|0030) refers to the physical distance in the patient between the center of each pixel, ranging from 0.461 to 0.977 mm. We normalise the lung CT images by resampling to $1 \times 1mm$ in 2D nodule segmentation researches.

There are some works that utilised spline interpolation method to perform resampling, but we found that the maximum and minimum pixel values may change after the operation, resulting in a different contrast compare to the original image. The range of pixel values varies from slice to slice, so we argue that it is not an optimal method. Alternatively, we performed the

resampling by simply resizing the images based on the pixel spacing attributes. The range of the pixel values is preserved in [-1024, 800], the same amount before the process.

After resampling, the images became smaller and different in size. It is sub-optimal to train a generative adversarial model with different size of images, and zero padding was implemented to obtain the original size of images while preserving the resampled lung CT region. The same process was performed on corresponding semantic label maps. In the end, a size of 512×512 ground truth image and semantic label map are acquired, shown as Figure3.8.



(a) Ground Truth (b) Lung semantic label

Figure 3.8: Original and resampled lung CT images.

3.2 Experiments

The experiments can be divided into two parts, generation of lung CT images and suitability for data augmentation in Section 3.2.1 and 3.2.2 respectively. For the first experiment, the main goal is to evaluate the synthesized images. In order to meet that goal, a series of trials are conducted, including fidelity evaluation, model comparison and manipulation test. The second experiment is to test the usability of proposed model as data augmentation method, and the training strategy is different from the previous experiment.

3.2.1 Generation of lung CT images

In this research, we trained the model with 2D lung nodule slices. To train and validate the model properly in medical image synthesis, we split the dataset into train and test sets in patient-based. In other words, the patients in the test set will be only seen in the test set, and so does the train set.

As previously mentioned, we selected 20% of the slices from each scan, including nodule and non-nodule slices. The selection resulted in 49992 images, and we further split it into train and test set, 42952 and 5616 respectively. The train set contained the selected slices from patient number 1 to 900, which is 806 patients in total with 2 scans unavailable (patient number 238 and 585) and 92 scans thickness greater than 2.5 mm. The test set was consisted of slices from patient number 901 to 1012, 90 patients in total with 22 scnas thickness greater than 2.5 mm.

3.2.1.1 Fidelity evaluation

The experiment was designed to answer the following research question: how perceptual realistic are the generated images. The goal of this experiment is to train the SPADE with the lung CT images and generated semantic label map and validate the quality of synthesis results.

We trained and validated the model by the proposed train and test set. Through the training process, the ground truth images were used as style guidance for which the variational autoencoder (VAE) will learn to extract meaningful features for the generator. Combined with the semantic label maps, the semantic information will be passed through the SPADE module to influence the synthesis process. The synthesis results will be concatenated with semantic label maps and fed into the discriminator to learn to distinguish the real and fake samples, as shown in Figure2.2. Following the data format of literature [10], the ground truth images were stored as colour images with three channels in JPEG type, while the semantic label maps were stored as grey images in PNG type.

It is challenging to select the metrics for the evaluation of medical image synthesis. Researches have utilised traditional shallow reference metrics such as mean-squared error (MSE), peak signal-to-nose ratio (PSNR), structural similarity index (SSIM) and multi-scale SSIM (MS-SSIM). However, these measures do not correspond to the visual quality of the images directly [8]. For example, direct optimisation of pixel-wise loss function will result in a blurred result. Therefore, it is not optimal to utilise them alone.

An alternative way is to utilise the down stream tasks such as a segmentation network to validate the quality of synthesis samples. A well-trained segmentation network should be capable of determining the discriminative features of the target tissues. Hence, if the synthesis samples are realistic, the network should be able to segment it accurately. On the other hand, qualifying by segmentation network is also meaningful for semantic image synthesis, since it can be interpreted as the style transfer loss between the semantic labels and the synthesis images. While the semantic label stands for the source domain, the synthesis process transfers the style from the semantic label to the target domain, the ground truth images. As the segmentation networks transfer reversely from the target domain to source domain, the semantic label should be the same as the original one if the translations are done perfectly, and the difference between them refers to style loss, or cycle-consistency loss [51]. Therefore, for a semantic image synthesis model, cycle-consistency loss can be interpreted as a quantitative measure of how precise the generative model follows the semantic order by users. Many state-of-the-art researches [10][11][30] have adopted this method as one of the quantitative metrics.

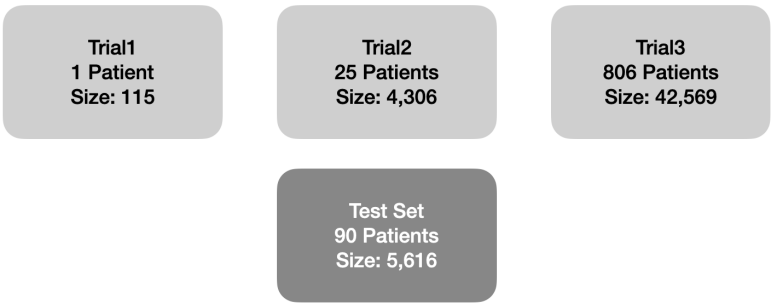


Figure 3.9: To validate the evaluation metrics, 3 experiments were designed with different train sets, and evaluated by the same test set. SPADE is trained in all trials.

The measures mentioned above provide alternative methods for quantitative measurement, but they do not directly represent the perceptual quality. Recently, some works have been focused on developing evaluation metrics that can directly represent the human perceptual judgement. [52] has proposed learned perceptual image path similarity (LPIPS) based on the features learned by the deep learning networks. Compare to previous metrics, LPIPS outperforms the others in terms of agreement with human judgement, and it has been adopted in MedGAN [53] as quantitative metric for medical image transfer task. Another popular choice for evaluating perceptual quality of the images is Fréchet Inception Distance (FID) [54]. The

FID compares the distribution of generated samples with the distribution of real images, and is used to evaluate the quality of generated images in many GAN-based researches. However, the validity of FID for medical imaging remains to be explored [8].

As an exploration research, we want to evaluate how perceptually realistic the generated sample is compared with the ground truth. In order to achieve that goal, it is necessary to know which metric is valid for evaluating synthetic medical images. Since not much researches have synthesized the full lung CT images, we think it is worthy to conduct such comparison in our study. To do this, we additionally trained the model twice with different train sets. One was trained with only one patient with full slices from scan, and the other was trained with twenty five patients also with full slices. We denoted the first one as Trial1, and the latter as Trial2. The model trained with the original train set was denoted as Trial3, which had the largest train set compared to the others, as shown in Figure 3.9. By expectation, Trial3 should outperform Trial2, and Trial1 should have the lowest performance. We observed the synthesis results perceptually and compared them with the metrics. We expected that the difference in perceptual similarity should be reflected by the difference of metrics.

After we validate the evaluation metrics, we then proceed to evaluate the performance of SPADE in synthesizing lung CT images, following the standard metrics of semantic image synthesis [11] [30]. FID, mean intersection-over-union (mIoU) and pixel accuracy were utilised, and a Nested U-Net [34] capable of multi-classes segmentation was trained additionally to measure the mIoU and accuracy.

3.2.1.2 Model Comparison

As a comparison baseline, we chose pix2pixHD model, the precedent of SPADE, to train and compare the performance. We also follow the paper [11] to evaluate qualitatively the ability of multi-modal synthesis. To measure the variation in the multi-modal generation, the same semantic label will be synthesized 20 times, and MS-SSIM and LPIPS were used to evaluate the diversity between images generated from the same semantic label, shown as Figure3.10. For our purpose, the multi-modal synthesis is also an important attribute that we expect the simulator would have, and conducting such test would render a quantitative result of how diverse samples the models could generate. pix2pixHD was adopted as the baseline model to compare with.

3.2.1.3 Manipulation test

With semantic image synthesis, the users are able to generate realistic samples by editing the semantic label map. In this experiment, we mimicked the samples by changing the location of nodules and see how the location affected the nodule synthesis, including some extreme cases such as nodules outside the body or crossing through other tissues, shown in Figure3.11. The goal of the experiment is to answer the research question: to what degree can we influence the outcome by user input. By doing this, we could simulate the user editing rarely seen situations and observe how the model react to it. We utilised Trial2 and Trial3 to perform such tests, selecting the slice which has a visible nodule in the test set and edited the semantic label. We expected that the Trial3 should render more realistic samples than Trial2 since it had seen more samples. Since there was no ground truth available, we compared the results by perceptual check only, zooming in the nodule to see the detailed texture. Besides this, we also synthesized the nodule in an expanded and shrunken version.

3.2.2 Suitability for data augmentation

In this experiment, we explored the effectiveness of SPADE as a data augmentation method for nodule segmentation. To do that, we generated a new dataset with only nodule slices included. The nodules annotated by at least three radiologists were used in this experiment, ending up

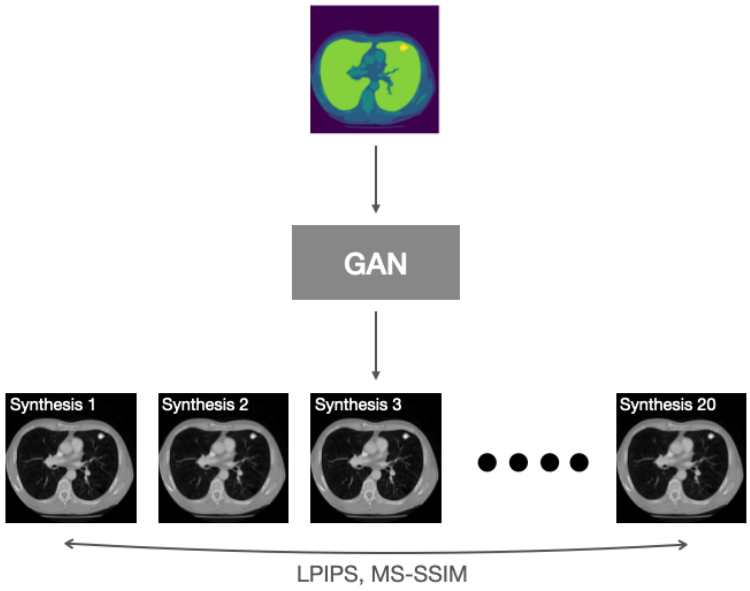


Figure 3.10: The semantic label was fed into the model and generated 20 synthetic results. The ability of multi-modal synthesis was quantitatively measured by calculating LPIPS and MS-SSIM between the synthetic results pair-wisely.

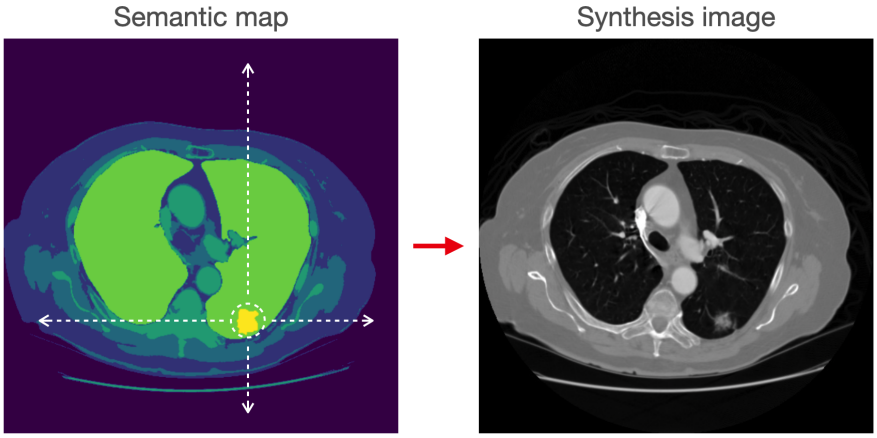


Figure 3.11: By editing the semantic label, the nodule is shifted horizontally and vertically. Through the manipulation test, synthesis of the nodule in different location and size can be observed.

with 7641 slices. The data was split into train and test set, 5936 and 1705 respectively. See Figure3.12.

The dataset was split into 5 subsets based on patients. One of the subset is used as test set, while the others are used as train set. Besides, 10% of the train set data was selected randomly as validation set during the training process. For data augmentation, we followed the experiment steps of [18], training the SPADE with the train set and synthesizing based on label from the same set., as shown in Figure3.13. The generated samples were then combined with the original train set to form the augmented train set, and the amount of training data was doubled. The Nested U-Net was utilised as the 2D nodule segmentation network. In each trial, we trained the model with two different sets, original train set and the augmented set. The two

trained models were compared based on the same test set. The segmentation performance was evaluated quantitatively by dice coefficient, sensitivity and positive predictive value.

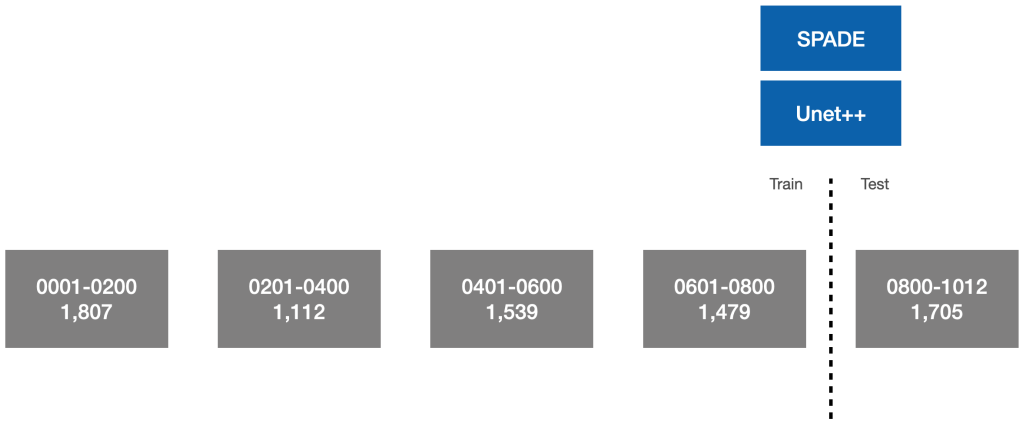


Figure 3.12: For nodule segmentation training, we utilised the nodules which were annotated by at least three radiologists and generated a new dataset to train SPADE and Nested U-Net.

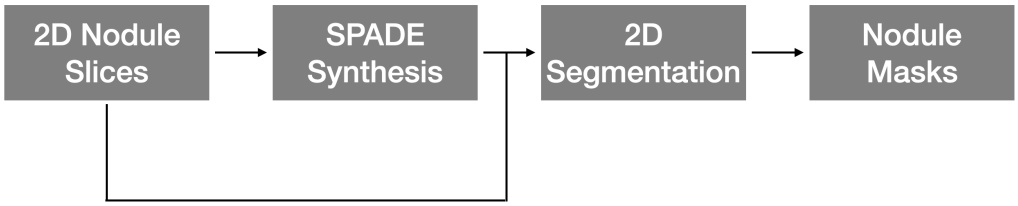


Figure 3.13: Synthetic CT images are generated, and then are combined with the original images to form the augmented train set for training the segmentation network.

3.3 Implementation detail

The GauGAN model was trained with 50 epochs and a batch size of 8. The spectral normalization [55] was applied in both generator and discriminator. The learning rates for the generator and discriminator are 0.0001 and 0.0004, following the setting of original literature. ADAM optimizer [56] was applied with $\beta_1 = 0$ and $\beta_2 = 0.9$. The two time-scale update rule (TTUR) [54] was applied for training due to the performance boost shown in the literature. We linearly decayed the learning rate to 0 from epoch 25 to 50. The images size was 512×512. Two NVIDIA A40 GPUs were used for training.

The Nested U-Net was trained with 150 epochs and batch size of 10. Adam optimizer was utilised and the multi-step learning rate was applied with $\gamma = 0.9$. Two decay milestones were set at epoch 30 and 75, following the default setting. The deep supervision [57], was enabled and the image size was 512×512 after resampling.

4 Results

In this chapter, the results of the experiments are shown. The first part is the generation of lung CT images (4.1) for which we examine the quality of synthetic images. Following sections are included in this part: fidelity evaluation (4.1.1), model comparison (4.1.2) and manipulation test (4.1.3). The second part is the 2D nodule segmentation. In this part, we evaluate the performance of SPADE as a data augmentation method (4.2), and the results are shown.

4.1 Generation of lung CT images

4.1.1 Fidelity evaluation

We started with qualitative comparisons. Figure 4.1 shows the comparison of real and synthetic CT images. The ground truth images are compared with the results of Trial1, Trial2 and Trial3. For synthesizing body, soft tissues and high dense tissues, the results of Trial3 are the most realistic due to the texture of tissues, while artefacts are visible in Trial1 and Trial2. We also found that the model learn to synthesize trachea without additional label. The trachea region, also known as windpipe, was not synthesized in Trial1, and was partially synthesized in Trial2, for which some of the trachea were completed and some of them were unclear. In Trial3, the trachea region was generated in the same position with ground truth. As for the bronchus, the airway in the respiratory system, was synthesized with visible artefacts in Trial1 and Trial2. In Trial3, it was synthesized without clear artefact based on visual examination without domain expertise.

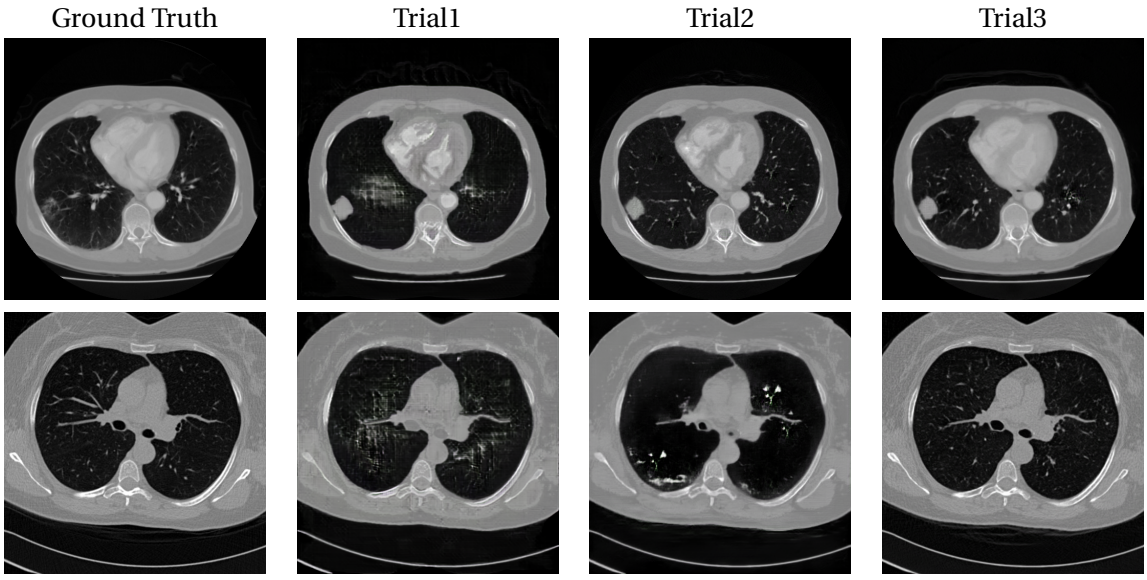


Figure 4.1: Visual comparison of semantic image synthesis results. SPADE is utilised in all trials. Trial3 with the largest train set, renders the most realistic CT images compared to the others.

For nodule synthesis, we compared the results of Trial2 and Trial3 with the ground truth shown as Figure 4.2. The results of Trial1 were not included in the comparison due to the quality. Compared to Trial3, Trial2 generated nodules with blurred boundaries and textures. Trial3 generally outperforms the others by visual examination, and we were able to utilise SPADE to mimic lung CT images from upper to lower lung. The comparison of synthetic nodules between trials are shown in Figure 4.3. Additional qualitative comparisons between trials are shown in Appendix A.

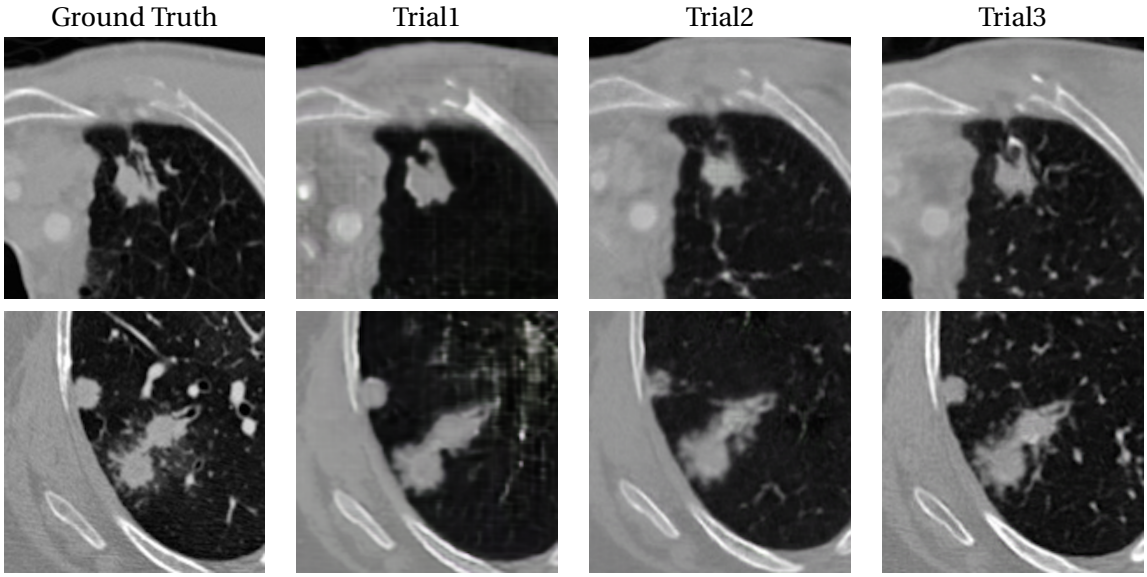


Figure 4.2: Visual comparison of nodule synthesis results between trials.

For quantitative comparisons, Trial3 is the most realistic, and Trial1 is the worst regarding most of the metrics. As shown in Table 4.1, Trial3 outperforms the others by a large margin in FID, and Trial2 is generally better than Trial1 except in MSE. For pixel wise metrics, Trial3 achieves the largest margin in MSE, while in PSNR the performance is slightly better than the others. For similarity, the difference between these trials is small in MS-SSIM, compared to the results in SSIM. As for perceptual metrics, the largest margin is achieved in FID, while in LPIPS the difference in performance is less but still aligned with the perceptual judgement.

	MSE ↓	**PSNR ↑**	**SSIM ↑**	**MS-SSIM ↑**	**FID ↓**	**LPIPS ↓**
Trial1	226.12 ± 89.18	24.88 ± 1.60	0.65 ± 0.09	0.84 ± 0.04	64.89	0.26 ± 0.08
Trial2	259.78 ± 432.20	25.01 ± 2.29	0.70 ± 0.11	0.85 ± 0.06	28.43	0.22 ± 0.10
Trial3	160.04 ± 69.05	26.48 ± 1.85	0.73 ± 0.08	0.86 ± 0.04	8.23	0.13 ± 0.03

Table 4.1: Quantitative comparison of semantic image synthesis result. Trial3 outperforms the others. For MSE, FID and LPIPS, lower is better. For PSNR, SSIM, MS-SSIM, higher is better.

We also evaluate the performance by the metrics which are common in the literature of semantic image synthesis, and the results are shown in Table 4.2. FID is used to evaluate the synthetic quality, while mIoU and pixel accuracy are used to evaluate semantic segmentation, which refers to the degree that synthesis process followed semantic order. An additional segmentation network is trained and used to segment the synthetic results, and the mIoU and pixel accuracy are calculated based on the results. Overall, Trial3 shows a better performance in semantic image synthesis, while the depicted content in Trial2 deviates the most from the input segmentation mask.

The analysis of segmentation for each class is shown in Table 4.3. Among each class, the highest IoU and pixel accuracy are achieved by Class0 and Class4, the background and the body, through all experiments. Class5, the nodule class, has the lowest segmentation performance, and Trial2 achieves the lowest IoU in Class5 compared to the others. Trial3 does not surpass Trial1 in IoU of nodule.

	mIoU ↑	FID ↓	accu ↑
Trial1	0.74	64.89	0.93
Trial2	0.71	28.43	0.92
Trial3	0.76	8.23	0.94

Table 4.2: Trial3 outperforms the others in semantic segmentation (mIoU and accu) and FID. For the mIoU and accu, higher is better. For FID, lower is better.

	Trial1		**Trial2**		**Trial3**	
	IoU	**accu**	**IoU**	**accu**	**IoU**	**accu**
Class0	0.99	0.99	0.98	0.99	0.99	0.99
Class1	0.83	0.94	0.83	0.94	0.87	0.95
Class2	0.83	0.94	0.82	0.93	0.86	0.94
Class3	0.83	0.98	0.82	0.98	0.85	0.99
Class4	0.99	0.99	0.97	0.99	0.99	0.99
Class5	**0.6632**	0.99	**0.6061**	0.99	**0.6553**	0.99

Table 4.3: The analysis of segmentation for each class. Class0 is the background. Class5 the nodule class achieves the lowest IoU among all classes.

4.1.2 Model comparison

In this section, we would like to utilise another semantic image synthesis model to compare with. pix2pixHD, the precedent of SPADE, is used and trained with same implementation detail. For comparison, we consider not only the capability of semantic image synthesis, but also the ability of multimodal synthesis due to the project goal. In Table 4.4, the comparison of semantic image synthesis is showed. pix2pixHD generates synthetic CT images with better quality than SPADE. As for semantic segmentation, pix2pixHD is slightly better than SPADE in mIoU and accu.

In Table 4.5, the analysis of segmentation for two models are shown. pix2pixHD performs better than SPADE in IoU of Class5 and achieves equally in the others. The synthetic nodules from pix2pixHD and SPADE are shown in Figure 4.3. The qualitative comparison of lung synthesis is shown in Appendix A.

	mIoU ↑	FID ↓	accu ↑
pix2pixHD	0.7595	7.72	0.9424
SPADE	0.7586	8.23	0.9400

Table 4.4: SPADE does not surpass pix2pixHD in the quality of semantic image synthesis.

	pix2pixHD		**SPADE**	
	IoU	**accu**	**IoU**	**accu**
Class0	0.99	0.99	0.99	0.99
Class1	0.87	0.95	0.87	0.95
Class2	0.86	0.94	0.86	0.95
Class3	0.85	0.99	0.85	0.99
Class4	0.99	0.99	0.99	0.99
Class5	**0.6625**	0.99	**0.6553**	0.99

Table 4.5: The analysis of segmentation for pix2pixHD and SPADE.

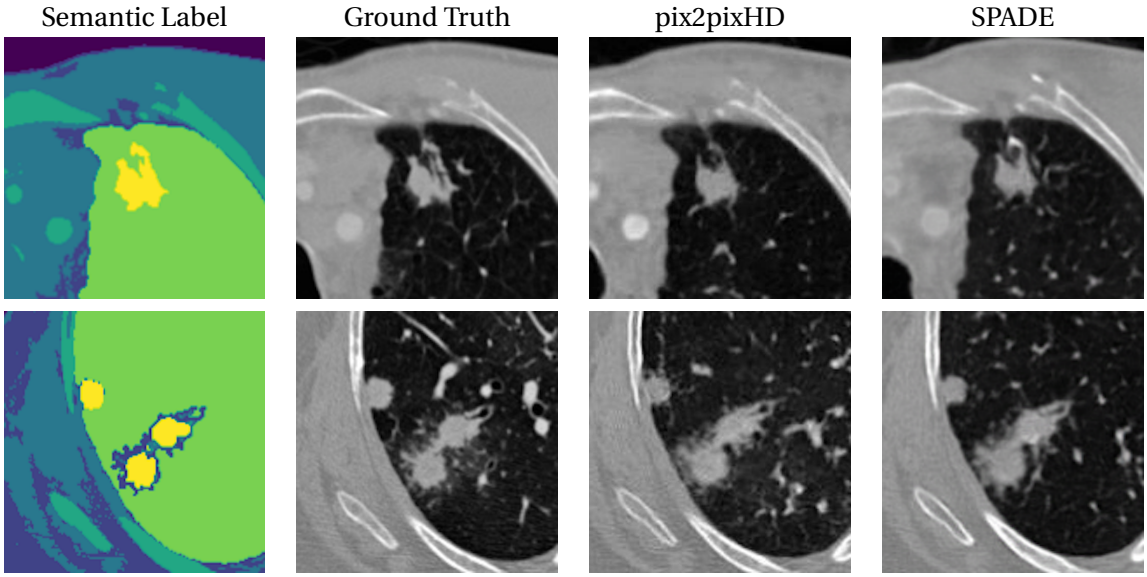


Figure 4.3: Qualitative comparison of nodule synthesis results generated by pix2pixHD and SPADE (Trial3).

In Table 4.6, the results of multimodal synthesis are showed. With same semantic input, pix2pixHD achieves 0 in LPIPS and 1 in MS-SSIM, indicating it could only synthesize identical images, while SPADE can achieve different outputs. Therefore, SPADE is better than pix2pixHD in multimodal synthesis.

The results of multimodal synthesis are shown in the following figures. In Figure 4.4, we show that the synthetic tissues can be synthesized differently. It can be the texture, contrast or the content. For example, the air bags in the chest are generated differently with same semantic input. The nodules can also be synthesized differently, as shown in Figure 4.5

	LPIPS ↑	**MS-SSIM ↓**
pix2pixHD	0	1.000
SPADE	0.022 ± 0.013	0.987 ± 0.007

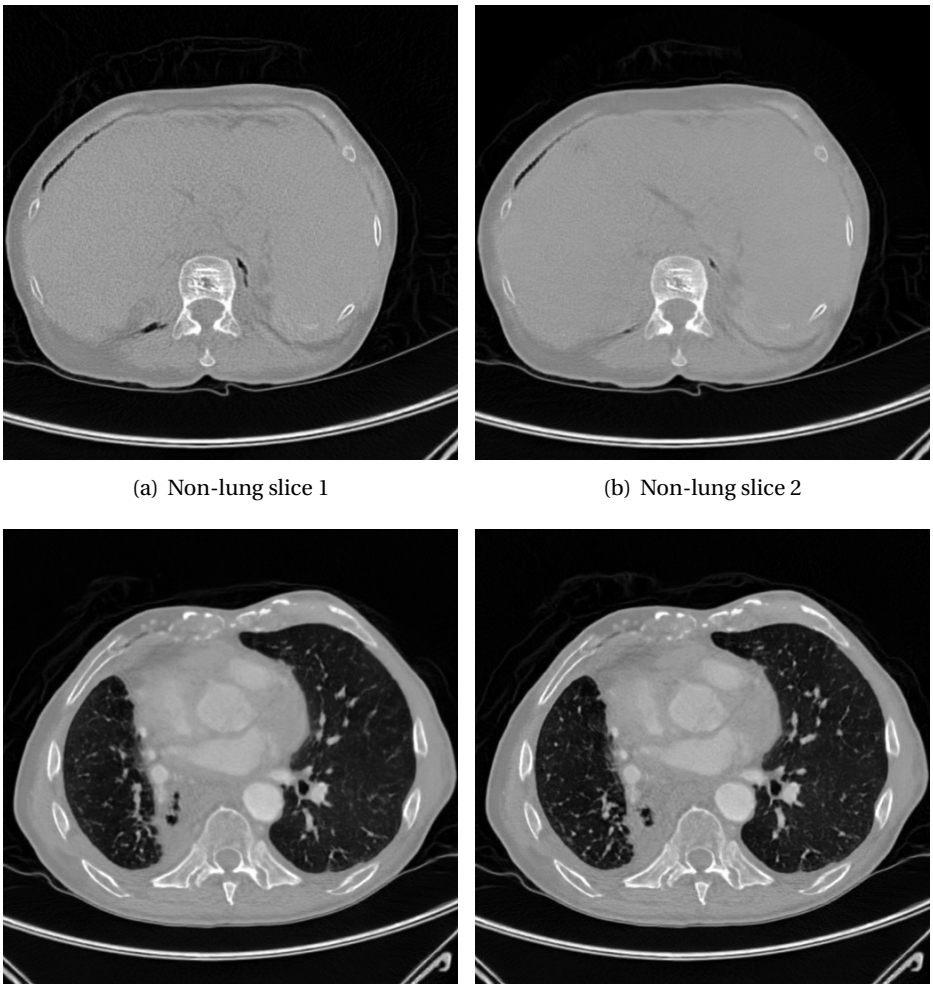
Table 4.6: The capability of multi-modal synthesis is evaluated quantitatively by pair-wise calculation between synthetic images. SPADE outperforms pix2pixHD in multimodal synthesis.

4.1.3 Manipulation test

In the manipulation test, the synthetic results was manipulating by semantic labels that nodule is relocated horizontally and vertically. The results are shown in Figure 4.6 and Figure 4.7 respectively. The nodules are not synthesized identically through the process in different position. Instead, the nodule characteristic such as texture would change based on the position and the neighbour tissues. We also resize the nodules, and visually examine the results shown in Figure 4.8. Even though the nodules are synthesized in the same position, the property of nodule would also change based on the size.

4.2 Suitability for data augmentation

The segment network trained by original data without augmentation is referred as baseline. The proposed method indicates train set is augmented by 20% synthetic images with horizontal flip operation, while traditional method represents the augmentation by 20% ground truth data with flip operation. The results are shown in Table 4.7. Compared to the baseline, the proposed data augmentation method improves the dice coefficient 0.7% and sensitivity 3.65%,



(a) Non-lung slice 1 (b) Non-lung slice 2

(c) Lung slice 1 (d) Lung slice 2

Figure 4.4: The capability of multimodal synthesis allows synthesizing slightly different outputs with same semantic label. The difference can be texture, content or the contrast of images. All the results are synthesized by SPADE.

while dropping 3.29% in PPV due to the increasing amount of false positive. The traditional method outperforms the proposed method in dice coefficient and PPV by 1.84% and 5.13% respectively, but in sensitivity the proposed method shows slightly advance with 0.78%.

	DSC (%)	**PPV** (%)	**SEN** (%)
Baseline	71.47	79.40	64.97
SPADE	72.17	76.11	68.62
Traditional	73.94	81.24	67.84

Table 4.7: Comparison of data augmentation method. SPADE as data augmentation method improves the dice coefficient and sensitivity, but it does not surpass the traditional approach.

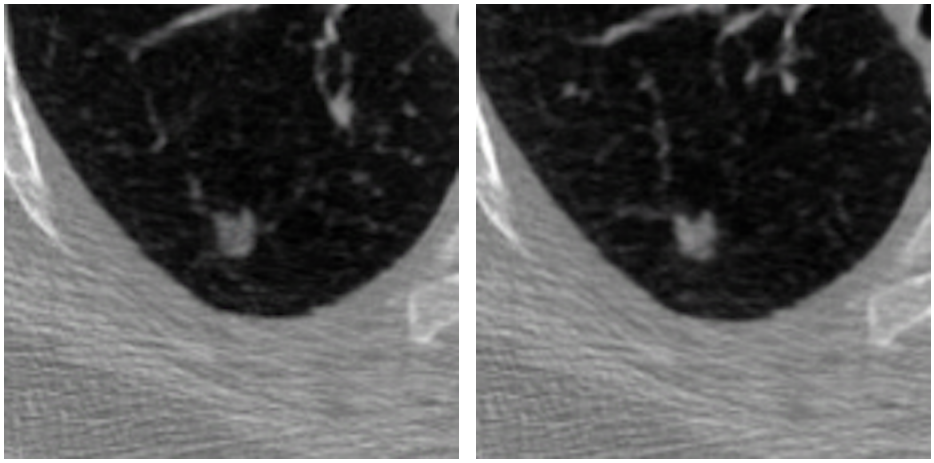


Figure 4.5: Nodules are synthesized differently due to multimodal synthesis.

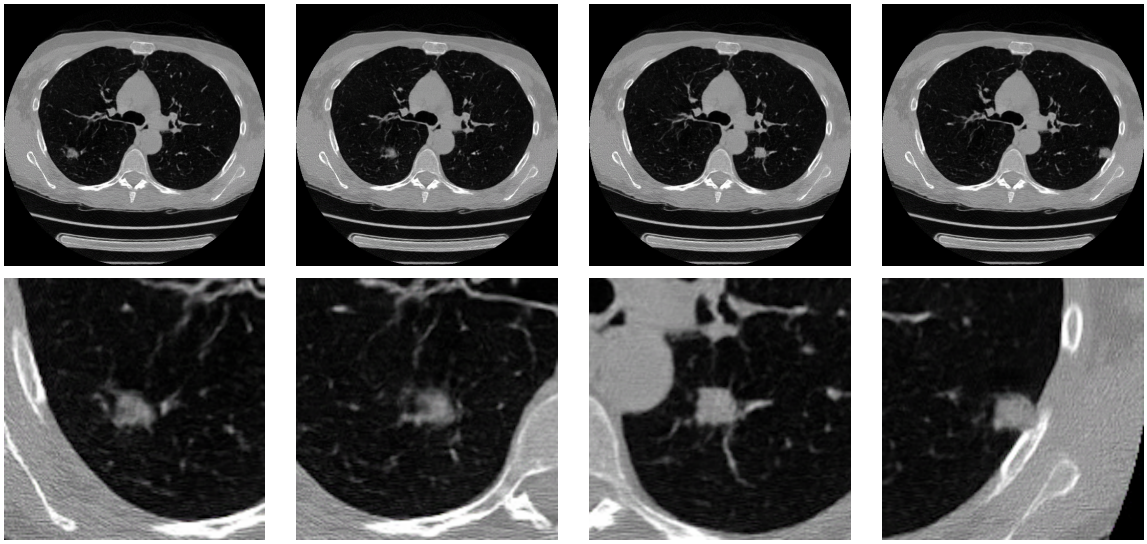


Figure 4.6: We relocate the nodules horizontally and synthesize based on the manipulated labels.

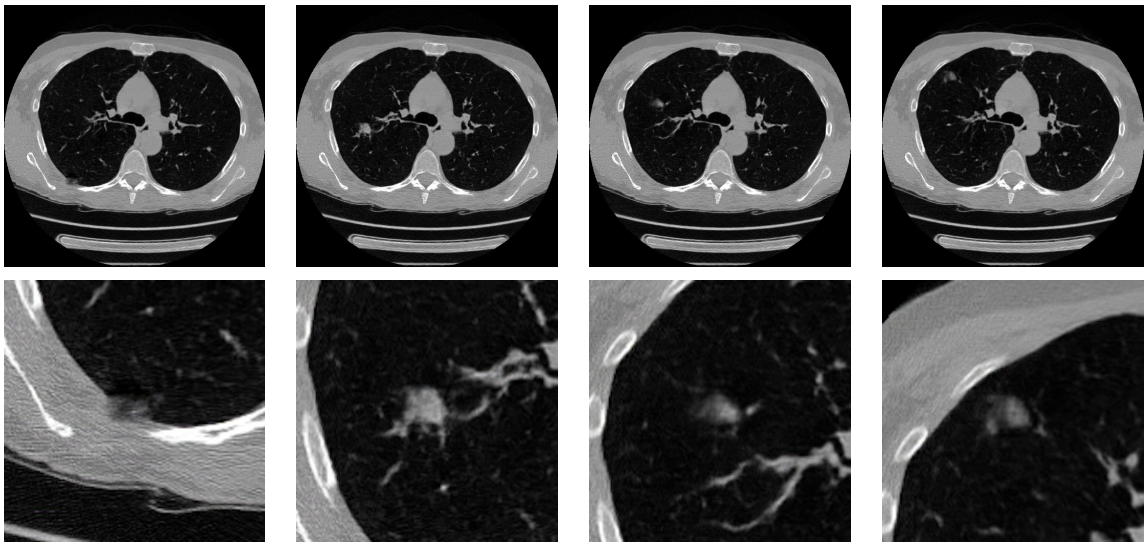


Figure 4.7: We relocate the nodules vertically and synthesize based on the manipulated labels.

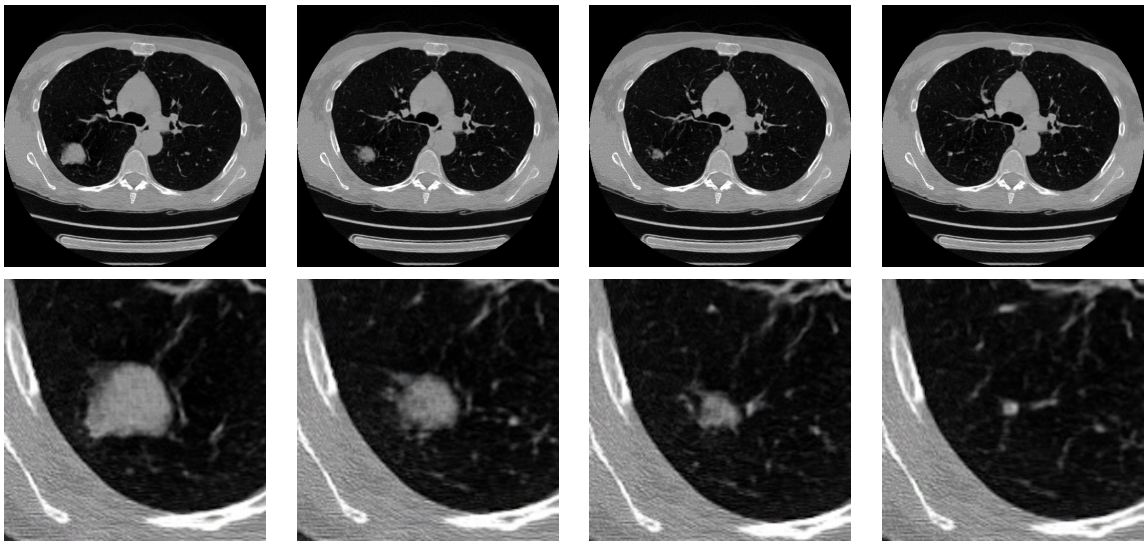


Figure 4.8: We also expand and shrink the nodule. From left to right, the nodules are manipulated in size by 2x, 1.5x, 1x and 0.5x.

5 Discussion

5.1 Generation of Lung CT Images

The first part of this section will be the interpretation for lung CT generation (5.1) covering fidelity evaluation (5.1.1), model comparison (5.1.2) and manipulation test (5.1.3). For the second part, the results of suitability for data augmentation (5.2) will be disccused.

5.1.1 Fidelity evaluation

In Table 4.1, most of the selected metrics align with visual examination, except Trial2 achieves the largest mean and deviation in MSE. It is because of mode collapses exist in the inference samples that the gray scale images were coloured. This indicates that SPADE could still suffer from mode collapse in semantic image synthesis. While Trial3 achieves the best results in MSE and PSNR, we argue that pixel-wise metrics are suboptimal for semantic image synthesis, since synthetic tissues could have different texture than ground truth but still remain photorealistic, just as the cases we have observed in multimodal synthesis test. Although MSE gives a distinguishable result and PSNR reflects the degree of blur, both could not correspond directly to the fidelity.

For the rest of the metrics that align with visual examination, SSIM and MS-SSIM do not render distinguishable results. In some cases, synthetic images with worse quality could achieve better scores in MS-SSIM than the good one. Therefore, both structural similarity metrics are also suboptimal for semantic image synthesis. For perceptual similarity metrics, the largest margin is achieved in FID that Trial3 is 7.8 times better than Trial1 and 3.4 times better than Trial2, given the most distinguishable metrics among the others. On the other hand, LPIPS renders a relative small margin compared to FID. By comparing the mean LPIPS values, Trial3 is 2 times better than Trial1 and 1.7 times better than Trial2. Besides, the distinction between Trial1 and Trial2 in LPIPS is not divisible enough as the visual examination. Therefore, compared to LPIPS, FID can reflect better the difference of quality for synthesis lung CT images.

In Table 4.2, along with FID, the common metrics for semantic image synthesis are applied for evaluating Trial1, Trial2 and Trial3. For semantic segmentation, Trial3 performs slightly better than Trial1 and Trial2, indicating that class 1 to 3 can be synthesized easily with small data and recognised by the segmentation network. The class 4, the lung class, is easily recognised even with obvious artifacts due to the clear boundary between dark and bright regions. The class 5, the nodule class, is the hardest part for the segmentation network to precisely locate due to the small area and partially solid boundary. The mode collapse in Trial2 is also reflected by semantic segmentation, that the intersection over union for each class decays due to the unnatural synthesis and the artifacts. Interestingly, the highest IoU of nodule class is achieved by Trial1. To interpret this, a qualitative evaluation is conducted on the synthetic nodules, shown as Figure [**?**]. The nodules from Trial1 are solid and uniform, for which are easy for the segmentation network to detect, but they are not as realistic as the results from Trial3. Therefore, the performance in IoU does not directly related to the fidelity, but it indicates that SPADE learns the discriminated features for the segmentation network.

We set up the experiments to train SPADE, and the amounts of train data for each experiments are largely different in order to have outcomes that can be judged easily without domain expertise. Trial3 with the most diverse and abundant training data is expected to outperform the rest, and we have verified it through visual examination. The metrics have been calculated and compared with visual examination. We found that most of the metrics align with perceptual observation, but only FID could fully reflect the level of difference in synthetic quality. Therefore, we think that FID as a perceptual similarity metric is valid for quantitative evaluation of

synthetic lung CT images. In latter experiment, we utilised FID along with mIoU and pixel accuracy, for which are common in semantic image synthesis literature. For semantic image synthesis, FID indicates the quality of images while the mIoU and pixel accuracy refers to the fidelity and the level of model following users' input, and the results also align with perceptual observation. Therefore, we have verified that the metrics of semantic image synthesis: FID, mIoU, Accu, are valid for the medical imaging. Future research of semantic image synthesis in the medical imaging could utilise these metrics for quantitative evaluation.

5.1.2 Model comparison

According to the literature [10], SPADE outperforms pix2pixHD in many datasets competition, but in Table 4.4, we find that it is not the case for the proposed dataset, and there are two possible reasons. First, medical imaging is different from regular data such as city scenes or natural images. Regular images contains colour objects in distant or closed views, while the CT images are 2D plane slices of the body depicted as grey level images. On one hand, the grey level images are easier for the model to learn the color appearance, while on the other hand, the model is required to learn the precise distribution of tissues. The synthetic medical samples from pix2pixHD are slightly better than those from SPADE in all metrics, and this indicates that the models which have been shown outstanding in normal images synthesis might not achieve the same degree of excellence in the medical imaging. The medical image synthesis could be a different challenge for generative adversarial networks.

Second, the modulation parameters learned by the segmentation masks make the generator synthesize repetitive texture in the class with large area. It maybe acceptable for natural images such as grass and sky to have similar texture across the class, while for medical imaging, those details are essential for quality evaluation. Compared to pix2pixHD, the samples from SPADE are locally realistic but globally monotonous especially in synthetic bronchi and vessel tissues, but the opinions from the domain experts are required to judge which one is qualitatively better. Besides, since the difference in FID between two models is small, the accuracy of FID in the medical images is needed to be verified.

In Table 4.6, SPADE demonstrates a level of multimodal synthesis that pix2pixHD is not capable of. The multimodal synthesis is due to the randomness by the image encoder. Compared to the regular dataset, the diversity of medical images that SPADE can render is much smaller, indicating that medical imaging is again a different challenge than normal computer vision task.

While SPADE is capable of multimodal synthesis in the medical imaging domain, it could be further improved by replacing the VAE structure. Literature has shown that replacing encoder structure with a different noise generator could improve the performance of multimodal synthesis, but at the same time decrease the synthetic quality. In other word, a tradeoff between quality and diversity exists for SPADE. That seems to be the obstacle for model optimisation, and this can be due to the VGG-based perceptual loss utilised along with the discriminator. As VGG network is pretrained on ImageNet, it balances the fidelity and diversity of synthetic samples, and methods relying on it would be constrained by the ImageNet domain, which does not include medical images, and the representation power of VGG. Even though we have demonstrated SPADE is capable of generating photorealistic lung CT images, whether such loss function is optimal for the medical imaging remains unclear. Besides, literature has shown that a better balance between multimodal synthesis and fidelity can be achieved by excluding perceptual loss. While it remains unclear whether this applies to the medical imaging, with recent progress on GAN's architecture designs and regularisation techniques, the actual necessity of the perceptual loss requires to be reviewed.

5.1.3 Manipulation test

Both vertical and horizontal nodule movement result in different synthetic outcomes. The difference can be attributes such as texture or subtlety, and it depends on the position and neighbouring tissues. The nodule can be solid texture when it is located in the middle of lung region, but become partially solid when it is close to the boundary. Currently we did not split the nodule class into multiple classes based on the attributes, so the model generally inference the samples based on what it has learn from the train set.

Interestingly, when we placed the nodule at the boundary of other tissues, it did not overlap the others abruptly. Instead, it merged into the surrounding tissues in order to comply the perceptual loss. This is a benefit for our application since it can prevent unnatural output when users accidentally overlap the nodule label to other tissues in the background. Besides, it seems that SPADE has learn some natural rules about lung nodule, that it has to be in the lung surrounding by air bag at least partially. When we placed the entire nodule inside the soft tissues, the SPADE generated the air bag between the nodule and soft tissues, indicating that SPADE has learned to manipulate the boundary in order to fit the manipulated object in with others. Spatially-adaptiveness seems to be an advantage of applying SPADE for user-guided image synthesis, but it would impact the segmentation performance that the segmented outline deviates from the given semantic label slightly. This also reflects the results in model comparison that SPADE achieves a lower mIoU compared to pix2pixHD.

By expanding and shrinking the semantic labels, the size of the nodules can be controlled, but artifacts become visible as the size increases. Since the big nodules are so limited in quantity, the synthesis results tend to be uniform and repetitive, making it less realistic.

5.2 Suitability for Data Augmentation

In Table 4.6, we demonstrated that the synthetic images could be used as augmented data to improve the segmentation network. Compared to the baseline, the dice coefficient and sensitivity are improved while the positive predictive values declines. Compared to the baseline and traditional approach, the best sensitivity is achieved by the proposed method. However, the proposed data augmentation approach does not outperform the traditional method on either dice coefficient and positive predictive values. Increasing the amount of augmented data would lead to overfitting and performance reduction for both traditional and proposed methods.

There are two possible reasons for such results. First, the synthetic nodules are not as diverse as the attributes presented in ground truth. Even though the segmentation network can learn nodules in the flipped location, most of them were synthesized as solid texture, following the majority of nodule types. As a result, the data augmented with flipped ground truth could render more performance improvement compared to the proposed method due to limitation in diversity. Furthermore, synthesizing images by flipping semantic labels might be a sub-optimal way to induce diversity. Second, the features learn by SPADE might not be realistic enough. SPADE is good at synthesizing large semantic region, but it does not perform well for fined details with small area. Compared to other tissues, lung nodules are usually small and the attributes can be varying, making it hard for SPADE to learn.

Several improvements can be made for the experiment. First, instead of flip operation, it is possible that other manipulation method can better improve diversity of synthetic nodules. We have shown that relocating the nodules could render different outputs due to spatially-awareness, so manipulation such as relocating and reshaping could be alternative operations to improve the degree of diversity. Second, classifying nodules into different classes based on the attributes could help SPADE to learn features separately and synthesize precisely. By separating nodule into different classes, users can control the synthetic attributes and improve the

class-imbalance problem between different majority and minority types, such as ground glass opacity (GGO). Third, a better training strategy can be proposed. The options include different manipulation operations and the quantity of augmented data. Optimising the training options could help improve SPADE-based data augmentation method.

6 Conclusion and Future Work

6.1 Conclusion

In this research, we studied the usability of semantic image synthesis in the medical images, for which we focused on lung CT images. In order to synthesize lung images, we proposed a preprocessing pipeline to generate semantic labels from CT images, which would then be used as the conditions for image generation process. As the nodule annotations were given, we used computer vision methods to divide the other tissues into 4 classes, resulting in 5 classes in total. The 5 classes semantic maps and the corresponding ground truth were used for training and synthesizing samples during inference. Our contribution includes:

1. Exploring semantic image synthesis methods to generate realistic lung CT images.
2. Proposing a preprocessing pipeline to generate semantic maps from lung CT images.
3. Proposing a SPADE-based data augmentation method.

In order to answer the three sub-questions and the main research question, the following experiments were designed and executed:

- For fidelity evaluation, the synthesis results were evaluated by visual examination and quantitative metrics. We also verified the validity of metrics in lung CT images by comparing perceptual quality with calculated measurement.
- For model comparison, two semantic image synthesis methods were trained and compared the qualities and the degree of multimodal synthesis.
- For manipulation test, in order to test the synthetic performance, the semantic mask of nodules were manipulated with different operations. We visually checked and analysed the outcome.
- As suitability for data augmentation, the synthetic images were used as augmented data to train the segmentation network. We compared the segmentation performance with and without the augmented data.

The first sub-question to be answered is: *"How perceptual realistic are the generated images?"* We evaluated the quality of synthesis qualitatively and quantitatively.The synthesis outcomes of a well-trained SPADE can be hardly distinguished from the ground truth images by visual inspections without domain expertise.

The second sub-question is: *"To what degree can we influence the outcome by user input?"* We narrowed down the region of interest to the nodules, and demonstrated that the nodules can be edited freely. However, the degree of fidelity will be maintained only if the user editing follows the anatomy. Therefore, as long as the semantic maps comply anthropotomy, it is likely that the synthesis results can be photorealistic even with manipulation on the nodules.

The third sub-question is: *"To what extent can the generated images be used for data augmentation?"* We trained the segmentation network with and without the additional synthesis data, and found that the performance improve slightly in dice coefficient and largely in sensitivity. However, compared to the traditional data augmentation method with flipping, the improvement by the proposed method does not surpass the traditional one.

These sub-questions helped to answer the main research question: **"To what extent can the semantic synthesis GANs be applied to lung CT images?"** Through the experiments, we have

shown that SPADE is an effective method for lung CT image generation due to its quality and manipulability. It is possible to further develop as CAL system for training junior professionals. Moreover, it can be used as data augmentation method for the medical imaging.

6.2 Future Works

Some recommendations for future work can be made. First, a perceptual study conducted by radiologists is recommended. There are some limitations with current measurements. The visual examination was done without the support of domain experts, so the synthetic tissues could be unrealistic from professional points of view. Besides, although the selected metrics are popular in the community, the validity are not verified in the medical imaging. In the experiment of fidelity evaluation, we have obtained discriminative outputs from several trials. The perceptual study can be conducted by presenting the these outcomes with the corresponding ground truths to radiologists for qualitative measurements. The qualitative evaluations by radiologists can be a direct assessment of the fidelity. Moreover, if the radiologists' judgements correlate with the quantitative measurements, the validation of the metrics for the medical images can be proven. A perceptual study can be seen in Appendix B, and we believe this will be helpful for future research in medical image synthesis.

Second, even though we have shown that the nodules can be synthesized realistically, the generated samples do present limitations. Since the nodules are not classified by the attributes, one of major limitation is that the features from all types could be present in a generated sample. It is also possible that the model will generalise the features after learning from an imbalanced dataset. This could eventually cause difficulty in interpreting visualisation for radiological residents. By classifying nodules into different classes based on the attributes such as malignancy score, the model can learn the most discriminative image features for each class, and the previous mentioned problems can be alleviated.

Third, CT scans are inherently volumetric, and the radiologists visualise the CT scans slice by slice during reading sessions. Hence, it is desirable for the CAL systems to synthesize axially consistent tissues, but the information of adjoined slices is not considered in synthesis process of the current method. Therefore, the possible development for CT scan generation can be either extending SPADE from 2D to 3D volumetric generation, or developing the loss function that maintain the continuity between connected slices.

Considering all the above-mentioned recommendations and detailed improvement in the discussion section, it is possible to develop a feasible and effective simulator of patient cases for educating radiological residents in diagnosis with advanced CAD systems. In the future, it is expected that the CAL systems can be one of the major instruments for the realisation of widespread implementation of lung cancer screening.

A Additional quantitative and qualitative results

In Table A.1, the additional segment performance for each trial are shown, including the ground truth.
In Figure A.1, we show additional synthesis results on the test sets with comparisons from Trial1, Trial2 and Trial3.
In Figure A.2 and A.3, we show the synthesis results from the SPADE on the lung and non-lung slices with comparison to those from pix2pixHD method.
In Figure A.5, we show that SPADE is able to synthesize lung CT images in different location of chest.

	Ground Truth		**Trial1**		**Trial2**		**Trial3**	
	IoU	**accu**	**IoU**	**accu**	**IoU**	**accu**	**IoU**	**accu**
Class0	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99
Class1	0.87	0.95	0.83	0.94	0.83	0.94	0.87	0.95
Class2	0.86	0.95	0.83	0.94	0.82	0.93	0.86	0.94
Class3	0.86	0.99	0.83	0.98	0.82	0.98	0.85	0.99
Class4	0.99	0.99	0.99	0.99	0.97	0.99	0.99	0.99
Class5	**0.6372**	0.99	**0.6632**	0.99	**0.6061**	0.99	**0.6553**	0.99

Table A.1: The analysis of segmentation for each class. Class0 is the background. Class5 the nodule class achieves the lowest IoU among all classes.

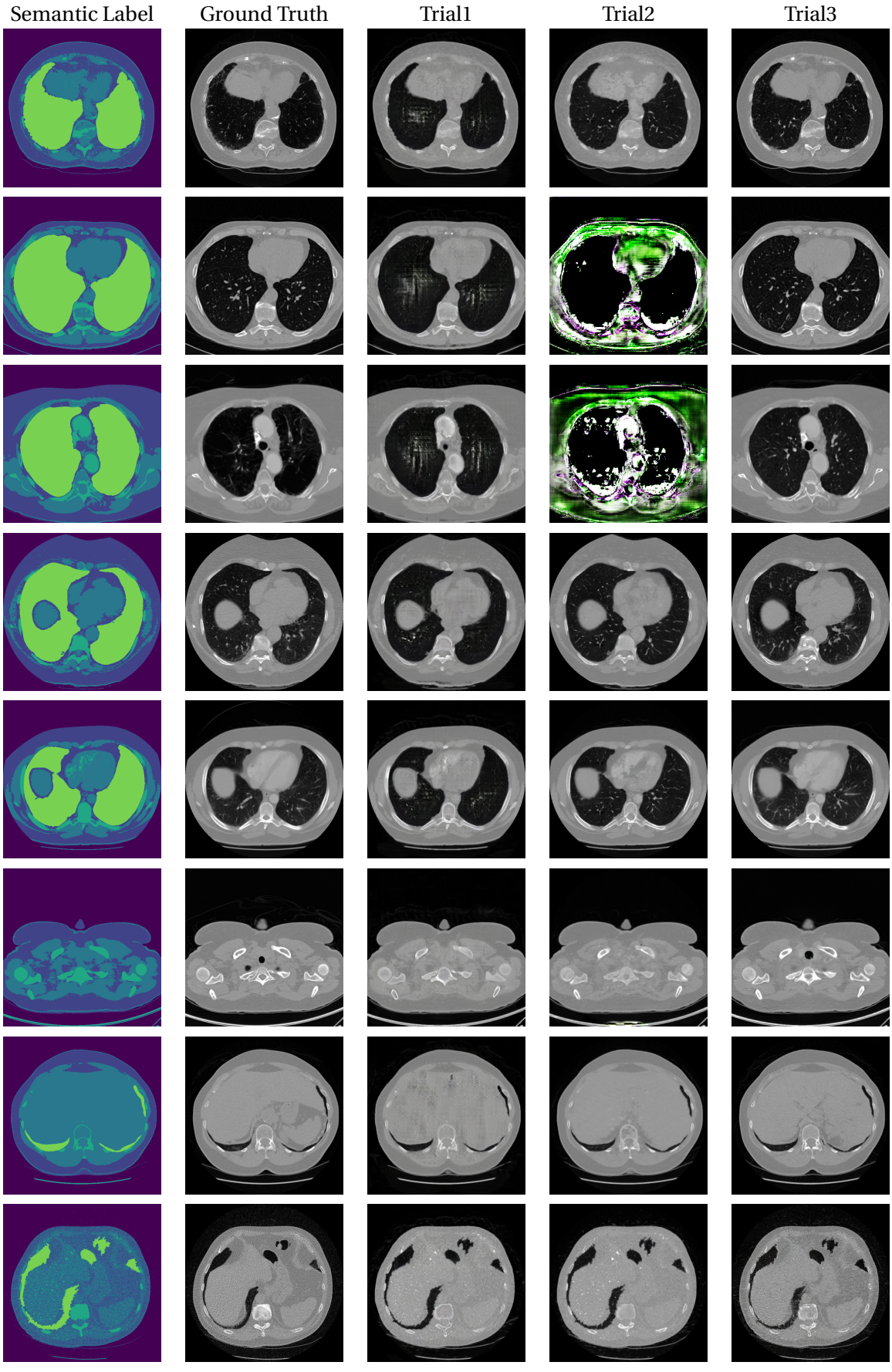


Figure A.1: Additional results with comparison to those from Trial1 and Trial2.

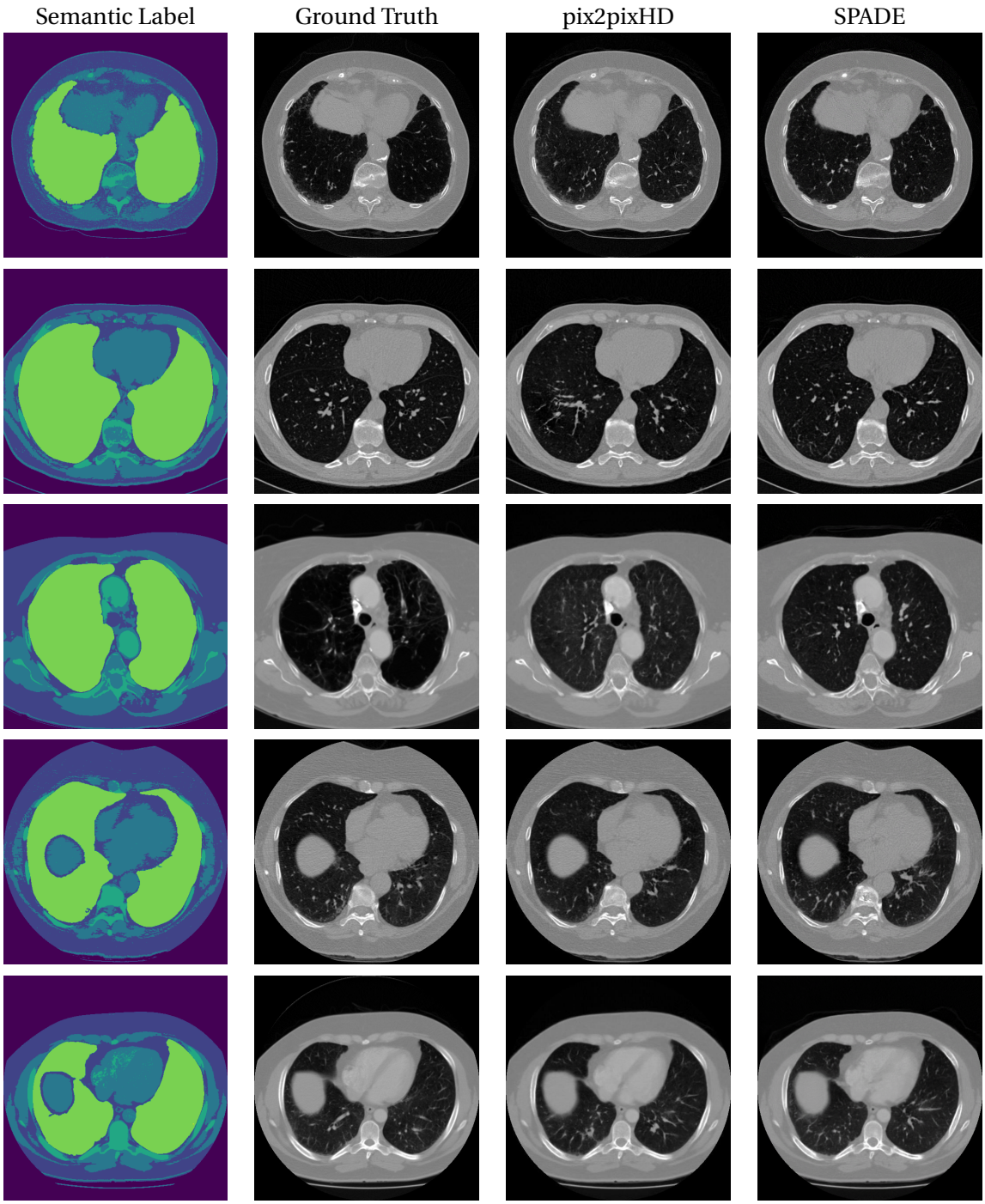


Figure A.2: Additional results with comparison to those from pix2pixHD and SPADE (Trial3) on the lung slices.

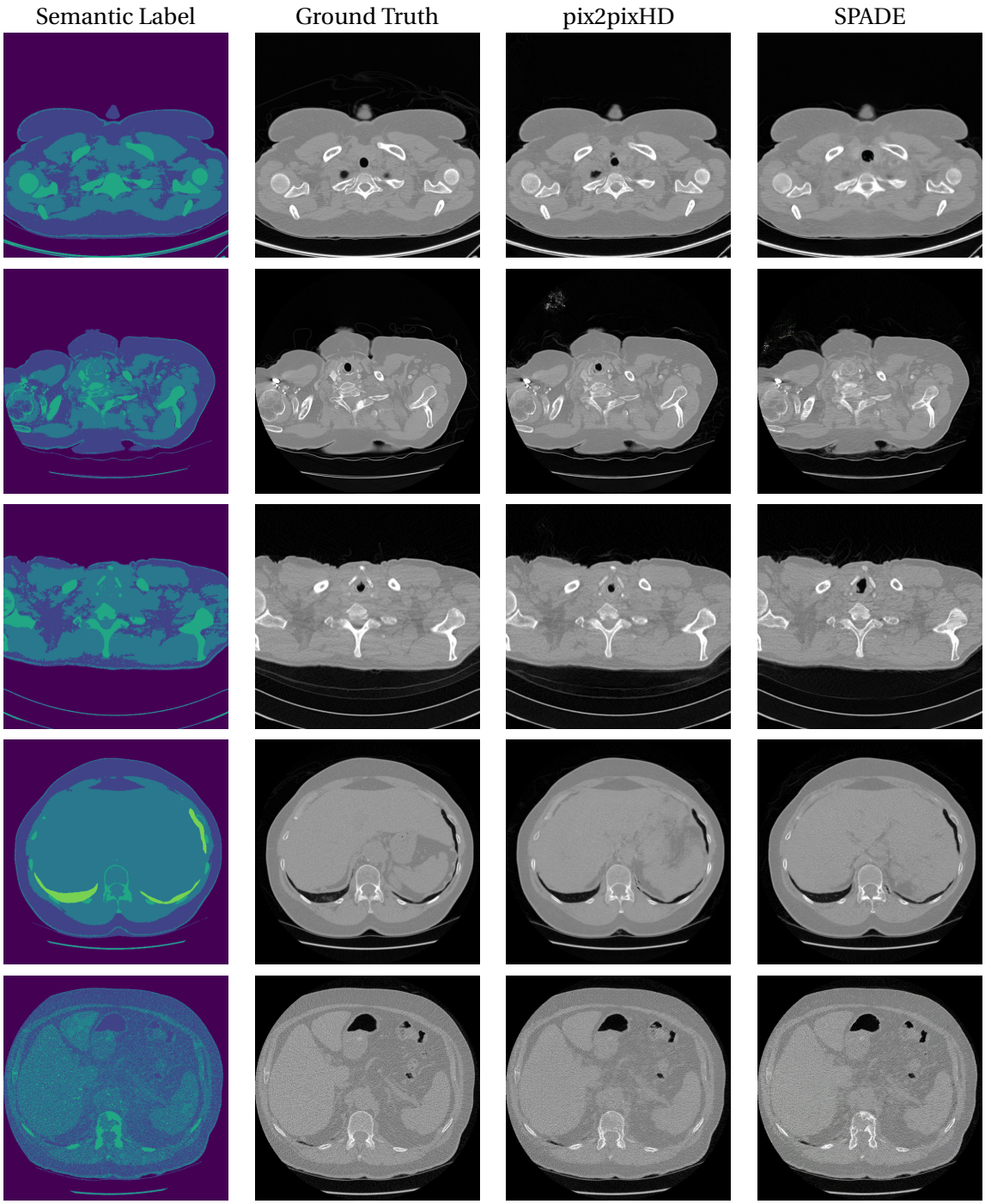


Figure A.3: Additional results with comparison to those from pix2pixHD and SPADE (Trial3) on the non-lung slices.

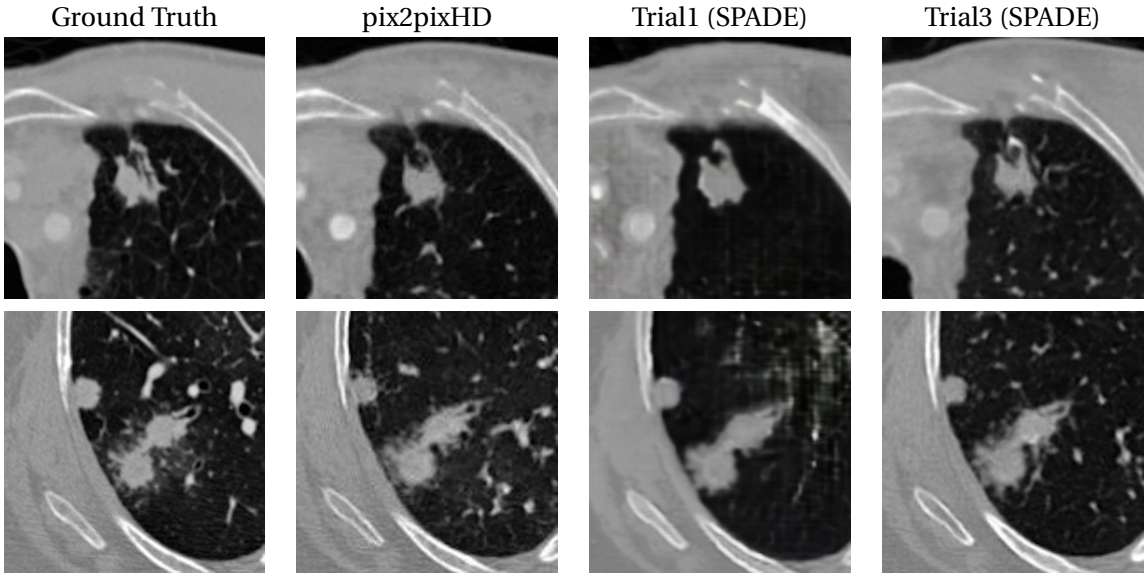


Figure A.4: Qualitative comparison of nodules from Ground Truth, pix2pixHD, Trial1 and Trial3. The IoUs (segment performance) for each nodule class are 0.6372, 0.6625, 0.6632 and 0.6553 respectively.

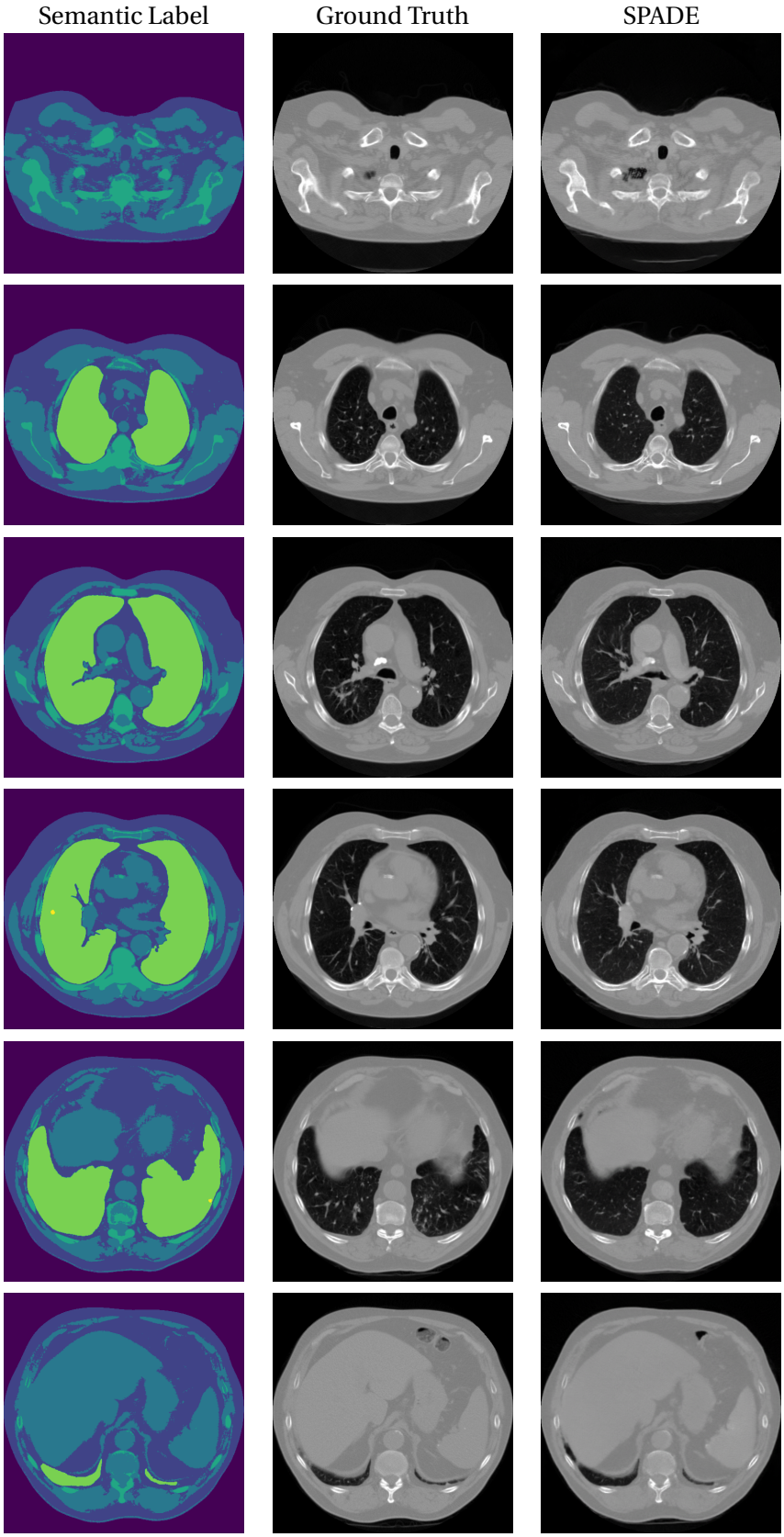


Figure A.5: A Full thorax CT images synthesis results from SPADE (Trial3).

B Future work: perceptual study and validation

B.1 Introduction

In the experiments of generation of lung CT images, the visual examination was conducted without the support from radiologists. The structure and the texture of the synthesized tissues may not be realistic from the professional points of view. Hence, considering the end users of the applications, we would like to have the radiologists to judge the fidelity of the synthesized images based on their domain expertise. The judgements can be used as not only an additional qualitative measurement, but also the validation of the quantitative measurements in applicability and accuracy.

B.2 Research questions

The purpose of this experiment is to answer the questions:

- Does the synthesis network render realistic CT images for the radiologists?
- Do quantitative image quality metrics align with the expert opinions?

B.3 Experiment designs

We adopt the methodology from previous work [53]. A series of experiments are designed for which we plan to have at least 5 radiologists examine the outcomes. There are 60 experiments, each containing the ground truth image and the synthesized outputs from Trial1, Trial2 and Trail3. In each experiment, the four images are presented in a randomized order. Participants are asked to choose one from the quartets which they think is real, and also evaluate the quality of each image using a 4-point score, with 1 being the most faked and 4 indicating the most realistic. All images are presented in 512×512 resolution. The same methodology can be applied for the comparison between those outcomes from SPADE and pix2pixHD for verifying the accuracy of quantitative image quality metrics.

The outcome of the questionnaires will be calculated based on the 4-point scores and real-image classifications. The mean value and standard deviation for each category represent the quality evaluation by experts, and the probability of the images classified as real indicates the degree of fidelity that the synthesis network can render. By comparing the outcomes of each category, we can correlate the expert opinions with quantitative image quality metrics.

For the infrastructure, we build the questionnaires using google form such that the participants can conduct the study in either the work place or home. In the following figures, we show the example of the questionnaires.

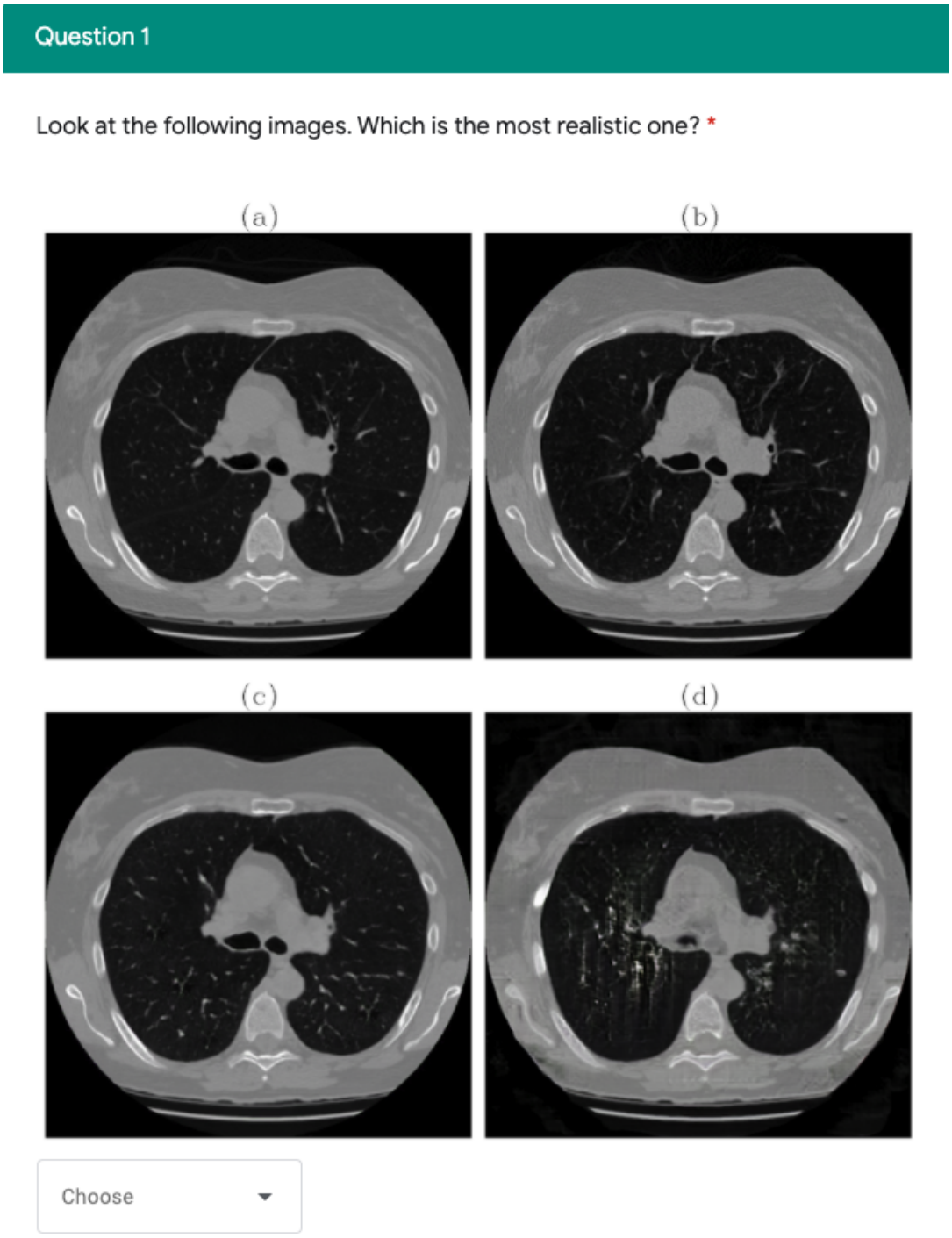


Figure B.1: The quartet is first presented for the participants to choose which one is most likely the real lung CT image.

Figure B.2: Each image is then presented in full resolution for the participants to rate the quality with 4-point score.

Bibliography

[1] Qiang Liu. Understanding the global cancer statistics 2018: implications for cancer control. *Science China Life Sciences*, 2019.

[2] M Malvezzi, G Carioli, P Bertuccio, P Boffetta, F Levi, C La Vecchia, and E2 Negri. European cancer mortality predictions for the year 2019 with focus on breast cancer. *Annals of Oncology*, 30(5):781–787, 2019.

[3] P David Mozley, Claus Bendtsen, Binsheng Zhao, Lawrence H Schwartz, Matthias Thorn, Yuanxin Rong, Luduan Zhang, Andrea Perrone, René Korn, and Andrew J Buckler. Measurement of tumor volumes improves recist-based response assessments in advanced lung cancer. *Translational oncology*, 5(1):19–25, 2012.

[4] Anand Devaraj, Bram van Ginneken, Arjun Nair, and David Baldwin. Use of volumetry for lung nodule management: theory and practice. *Radiology*, 284(3):630–644, 2017.

[5] Jan Hendrik Moltz, Lars Bornemann, Jan-Martin Kuhnigk, Volker Dicken, Elena Peitgen, Stephan Meier, Hendrik Bolte, Michael Fabel, Hans-Christian Bauknecht, Markus Hittinger, et al. Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in ct scans. *IEEE Journal of selected topics in signal processing*, 3(1):122–134, 2009.

[6] Jianrong Wu and Tianyi Qian. A survey of pulmonary nodule detection, segmentation and classification in computed tomography with deep learning techniques. *J. Med. Artif. Intell*, 2(8):1–12, 2019.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[8] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

[9] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020.

[10] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[11] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[13] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International conference on medical image computing and computer-assisted intervention*, pages 417–425. Springer, 2017.

[14] Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Išgum. Generative adversarial networks for noise reduction in low-dose ct. *IEEE transactions on medical imaging*, 36(12):2536–2545, 2017.

[15] Jaemin Son, Sang Jun Park, and Kyu-Hwan Jung. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. *arXiv preprint arXiv:1706.09318*,

2017.

[16] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.

[17] Maria JM Chuquicusma, Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 240–244. IEEE, 2018.

[18] Yulei Qin, Hao Zheng, Xiaolin Huang, Jie Yang, and Yue-Min Zhu. Pulmonary nodule segmentation with ct sample synthesis using adversarial networks. *Medical physics*, 46(3):1218–1229, 2019.

[19] Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K Kalra, Yi Zhang, Ling Sun, and Ge Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 37(6):1348–1357, 2018.

[20] Tony CW Mok and Albert CS Chung. Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks. In *International MICCAI Brainlesion Workshop*, pages 70–80. Springer, 2018.

[21] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[22] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[23] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. *arXiv preprint arXiv:1506.05751*, 2015.

[24] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2, 2014.

[25] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

[26] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[27] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.

[28] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[29] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018.

[30] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. Sesame: semantic editing of scenes by adding, manipulating or erasing objects. In *European Conference on Computer Vision*, pages 394–411. Springer, 2020.

[31] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *arXiv preprint*

arXiv:1910.06809, 2019.

[32] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[33] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[34] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.

[35] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. Gans for medical image analysis. *Artificial Intelligence in Medicine*, page 101938, 2020.

[36] Camilo Bermudez, Andrew J Plassard, Larry T Davis, Allen T Newton, Susan M Resnick, and Bennett A Landman. Learning implicit brain mri manifolds with deep learning. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105741L. International Society for Optics and Photonics, 2018.

[37] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[38] Ho Young Park, Hyun-Jin Bae, Gil-Sun Hong, Minjee Kim, JiHye Yun, Sungwon Park, Won Jung Chung, and NamKug Kim. Realistic high-resolution body computed tomography image synthesis by using progressive growing generative adversarial network: Visual turing test. *JMIR Medical Informatics*, 9(3):e23328, 2021.

[39] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[40] Mark Oldham, Jeffrey H Siewerdsen, Anil Shetty, and David A Jaffray. High resolution gel-dosimetry by optical-ct and mr scanning. *Medical physics*, 28(7):1436–1445, 2001.

[41] Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Išgum. Deep mr to ct synthesis using unpaired data. In *International workshop on simulation and synthesis in medical imaging*, pages 14–23. Springer, 2017.

[42] Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Gig S Mageras, Joseph O Deasy, and Harini Veeraraghavan. Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 777–785. Springer, 2018.

[43] Lei Bi, Jinman Kim, Ashnil Kumar, Dagan Feng, and Michael Fulham. Synthesis of positron emission tomography (pet) images via multi-channel generative adversarial networks (gans). In *molecular imaging, reconstruction and analysis of moving body organs, and stroke imaging and treatment*, pages 43–51. Springer, 2017.

[44] Yifan Jiang, Han Chen, Murray Loew, and Hanseok Ko. Covid-19 ct image synthesis with a conditional generative adversarial network. *IEEE Journal of Biomedical and Health Informatics*, 25(2):441–452, 2020.

[45] Covid-19, Feb 2020.

[46] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018.

[47] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[48] Scope and field of application.

[49] LLC Innolitics. Dicom standard browser.

[50] Menglu Liu, Junyu Dong, Xinghui Dong, Hui Yu, and Lin Qi. Segmentation of lung nodule in ct images based on mask r-cnn. In *2018 9th International Conference on Awareness Science and Technology (iCAST)*, pages 1–6. IEEE, 2018.

[51] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[53] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized Medical Imaging and Graphics*, 79:101684, 2020.

[54] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[55] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[56] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[57] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. PMLR, 2015.