RAM ROBOTICS AND MECHATRONICS

THE USABILITY OF GENERATIVE ADVERSARIAL NETWORKS FOR AUTOMATIC SEGMENTATION OF LUNG NODULES IN CT-IMAGES

S. (Sabien) van Elst

MSC ASSIGNMENT

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

May, 2021

028RaM2021
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 a highly prevailing disease and early detection and treatment is crucial to increase the likelihood of survival. In many lung cancer procedures, the segmentation of lung nodules in CT-images is an essential step. However, manual annotation of the nodules is a difficult and time-consuming task, which relies heavily on the experience of a radiologist. To assist radiologists in this process, computer-assisted segmentation systems could be a promising tool. The aim of this study was to explore the feasibility of applying generative adversarial networks (GANs) to automatically segment lung nodules from entire 2D CT-images.
The network architecture proposed in this thesis was designed to address three commonly occurring segmentation challenges: (i) variability in lung nodule appearance, (ii) class imbalance, and (iii) GAN training instability. The overall network followed the structure of a conditional image-to-image translation GAN, in which a U-Net was used as the generator network. To alleviate the aforementioned challenges, the following measures were adopted: atrous spatial pyramid pooling (ASPP) modules to reduce the influence of appearance variability, an additional Dice loss to counteract the class imbalance and a multiscale-L1 critic for adversarial training stability. The added value of each module to the generator network was evaluated in an ablation study. Additionally, the usefulness of GANs compared to the state-of-the-art segmentation network, the U-Net, was explored.
The results of the ablation study showed that the addition of ASPP modules to various locations of the generator network achieved inferior or comparable results to the results of the generator without ASPP modules. The Dice loss appeared to be an essential addition to produce accurate segmentation results for both the U-Net and the GAN. The GAN with a multiscale-L1 critic did show a stable training course, in contrast to more conventional GANs. However, introducing adversarial training by adding a critic did not improve the performance of the generator network alone. Overall, the GAN-based method obtained worse results than the U-Net, whereas the U-Net itself achieved excellent, state-of-the-art segmentation results compared to other works.
These results suggest that the Dice loss and the multiscale-L1 critic did facilitate in solving the segmentation challenges, whereas the ASPP modules were not of added value. Although the GAN with an additional Dice loss showed sufficient performance and comparable results to other GAN-based methods, it could not outperform the U-Net. In conclusion, this study showed that GANs can be applied for lung nodule segmentation in entire 2D CT-images, but the proposed GAN-architecture is not advantageous over the current state-of-the-art segmentation methods for this application. Future work could focus on expanding and balancing the dataset, using more suitable loss functions or extending the 2D models into 3D models.

Samenvatting

Longkanker is een veelvoorkomende ziekte en tijdige diagnose en behandeling ervan is cruciaal om de overlevingskans van patiënten te vergroten. De segmentatie van long nodules in CT-afbeeldingen is een essentiële stap voor de detectie en diagnose van longkanker. Handmatige annotatie van nodules is echter een ingewikkelde en tijdrovende taak, die sterk afhankelijk is van de ervaring van een radioloog. Computerondersteunde segmentatiesystemen zouden een veelbelovend hulpmiddel kunnen zijn om radiologen bij dit proces te assisteren. Het doel van deze studie was om te onderzoeken in welke mate Genenerative Adversarial Networks (GANs) gebruikt kunnen worden om automatisch long nodules uit volledige 2D CT-afbeeldingen te segmenteren.
De architectuur van het netwerk dat in deze studie voorgesteld wordt, is ontworpen om drie veel voorkomende segmentatie-uitdagingen aan te pakken: (i) variabiliteit in het uiterlijk van long nodules, (ii) ongebalanceerde klasseverdeling en (iii) GAN-trainingsinstabiliteit. Het netwerk heeft de structuur van een conditionele GAN die beelden van het ene domein naar het andere domein transleert. Hierin is een U-Net, een state-of-the-art segmentatie netwerk, gebruikt als generator netwerk. Om de bovengenoemde segmentatie-uitdagingen te verlichten, zijn de volgende modules geïmplementeerd in het netwerk: Atrous Spatial Pyramid Pooling (ASPP) modules om de invloed van de uiterlijke variabiliteit te verminderen, een Dice loss om het probleem van ongebalanceerde klassen tegen te gaan, en een multiscale-L1-critic om de stabiliteit van de GAN tijdens het trainen te vergroten. De toegevoegde waarde van iedere module aan het generator netwerk is beoordeeld met behulp van een ablatiestudie. Daarnaast is onderzocht of GANs ten opzichte van het U-Net toegevoegde waarde hebben voor het segmenteren van long nodules in volledige 2D CT-afbeeldingen.
De resultaten van de ablatiestudie toonden aan dat de toevoeging van ASPP-modules op verschillende locaties van het generator netwerk resulteerde in inferieure of vergelijkbare resultaten ten opzichte van de resultaten van het generator netwerk zonder ASPP-modules. De toevoeging van een Dice loss bleek essentieel te zijn om accurate segmentatieresultaten te verkrijgen voor zowel het U-Net als de GAN. In tegenstelling tot meer conventionele GANs vertoonde de voorgestelde GAN met een multiscale-L1-critic stabiele training. Echter, de toevoeging van een critic netwerk verbeterde de prestaties van het generator netwerk niet. Het GAN-model behaalde slechtere resultaten dan het U-Net. Het U-Net zelf behaalde superieure segmentatieresultaten in vergelijking met andere onderzoeken.
Deze resultaten suggereren dat een Dice loss en de multiscale-L1-critic positieve invloed hebben op het verminderen van de segmentatie-uitdagingen. De ASPP-modules bleken geen toegevoegde waarde te hebben. Hoewel de GAN met een Dice loss in staat was om long nodules te segmenteren en vergelijkbare resultaten vertoonde met andere GAN-methodes voor nodule segmentatie, was deze niet in staat om beter te presteren dan het U-Net. Alles overziend toonde deze studie aan dat GANs kunnen worden gebruikt voor long nodule segmentatie in volledige 2D CT-afbeeldingen, maar dat de voorgestelde GAN-architectuur hiervoor geen toegevoegde waarde heeft ten opzichte van de huidige state-of-the-art segmentatiemethoden. Toekomstige onderzoeken zouden zich kunnen richten op het uitbreiden en balanceren van de dataset, het gebruiken van geschiktere loss functies, of het uitbreiden van de 2D-netwerken naar 3D-netwerken.

Contents

1 Introduction

1.1 Motivation

Lung cancer is one of the highest prevailing types of cancer and remains the most common cause of cancer death worldwide. In 2020, 2.21 million cases of lung cancer were registered, accounting for 1.80 million deaths. [1] The mortality rate of lung cancer could be reduced by early detection and treatment of malignant tumors. In an early stage, the treatment of cancer is more effective, resulting in an increased likelihood of survival. [2] Screening programs have been set up that aim to detect cancer even before a patient develops symptoms. [3]
The most reliable and most commonly used imaging technique for the diagnosis of lung cancer is Computed Tomography (CT). [4] This non-invasive imaging technique constructs multiple 2D cross-sectional image slices which, when stacked together, generate a 3D image of the patient that shows the internal structures as well as any abnormalities. Lung cancer is characterized by the existence of small lesions, also known as lung nodules, that are visible on thoracic CT-images. [5] The detection and diagnosis of these nodules require examination of the CT-images by a radiologist, in which lung nodule segmentation has a crucial role. Nodule segmentation is the task of identifying the boundaries of nodules appearing on a thoracic CT-scan. Accurate nodule segmentation is an important step in many lung cancer procedures, such as diagnosis of tumor malignancy, detecting changes in nodule volume and monitoring the tumor response to therapy. [6] It assists the radiologist in diagnosing lung cancer and helps in training new radiologists. [7] However, nodule segmentation is often challenging due to variances in intensity and the heterogeneous nature of lung cancer. [4, 8] Moreover, as one CT-scan contains about 150 to 500 slices, it is a time-consuming and labor-intensive task. The accuracy and outcome rely heavily on the experience of the radiologist, making it subjective and prone to interobserver variability. [9] To assist the radiologist in accurately detecting and segmenting lung nodules, computer-aided diagnosis (CAD) systems could be a promising tool in the process of clinical decision making. [4, 6] CAD intends to provide a second, objective opinion and to reduce the interobserver variability by applying image processing, computer vision and machine-and deep learning techniques. [10, 11] The use of machine learning and deep learning for medical applications has progressed rapidly in the last years. [12] A relatively new type of deep learning method is the generative adversarial network (GAN). This image generation technique can be used to enlarge datasets, to improve image quality or to convert images from one domain to another. [13] However, when segmentation is viewed as the generation of an image containing an annotated area of interest, image generation techniques could, ideally, also be used for segmentation. In this research, the potential use of GANs for the automatic segmentation of lung nodules from thoracic CT-images is studied.

1.2 Background and related work

1.2.1 Deep learning for medical image analysis

With the increase in computing power and the emergence of big data, artificial intelligence (AI) has attained increased attention and wide application in recent years. One of the subfields of AI is machine learning. This includes techniques that, based on hand-crafted features, enable computers to acquire patterns from raw data and learn from experience. These mathematical models can be trained to generalize their learned expertise and deliver useful predictions on new, unseen data. [14] One of the areas of machine learning that has been evolving most recently is deep learning. [10] Deep learning is characterized by computation models that consist of multi-layered artificial neural networks. By automatically extracting features from raw data, these networks show improved performance in comparison to the more traditional machine learning methods. [15] The advancements in deep learning techniques have inspired researchers to incorporate deep learning to the field of medical image analysis. This was mainly triggered by the development of efficient convolutional neural networks (CNNs). [16, 17] CNN is a powerful method for image analysis by

learning useful representations and recognizing visual patterns from image data. It is a specific type of artificial neural network designed to preserve spatial relationships in the data. The input to a CNN is a grid-like structure, which can be a 2D or 3D image, and is fed through multiple layers of convolutions and pooling. In the convolution layers, filters are applied to automatically extract relevant features from the input by convolution operations. CNNs make use of parameter sharing, in which the same filter is applied over all locations of the input, producing a tensor of feature maps. This leads to a great reduction in the total amount of parameters that need to be learned, making them highly efficient. The pooling operations are applied to reduce the dimensions of the data, which decreases the computational cost and makes the network more robust. [14] CNNs are widely used for medical image analysis and show promising results in numerous tasks, such as classification, detection and segmentation. [18, 19]

1.2.2 Automatic lung nodule segmentation

Since manual annotation of nodule boundaries is time-consuming and subjective, researchers have been searching for accurate and reliable automated segmentation methods, which could be valuable for both clinical and research purposes.

Conventionally, automatic segmentation of lung nodules was performed using traditional image processing techniques [6], such as thresholding [20], morphological operations [21], region growing [22], and energy minimization based methods [23]. The CAD systems used for lung nodule detection generally consisted of two phases: nodule candidates detection and false positive reduction. The goal of the first phase was to detect nodules at a very high sensitivity based on the aforementioned traditional image processing techniques. In the second phase, the false positive nodules were reduced using a feature-based classifier. [6] These methods usually utilized constraints about nodule appearance or image intensity for segmentation. Later on, machine learning techniques became dominant for segmentation, which classified each pixel independently and required hand-crafted features designed by experts. [24, 25] However, these methods were not optimal for lung tumor segmentation, due to their limited ability to segment the challenging nodule types, their robustness on heterogeneous, low contrast CT-images and their automation level. [6]

Since 2012, the application of deep and efficient CNNs has rapidly become a prominent method for computer vision and medical image analysis. [16, 18] Although CNNs are typically applied for classification, numerous researchers have studied the use of CNNs to address the problem of medical image segmentation in the last years. [26] One of the first papers that covered the use of CNNs for medical image segmentation was published by Ciresan et al. [27] They employed a CNN for pixel-wise segmentation of biological membranes in electron microscopy images by predicting the label of each center pixel individually in a sliding-window fashion. [27] CNN-based methods were also adopted for segmenting lung nodules in CT-images, by predicting whether the center voxel of a patch belongs to the nodule or background. [28, 29] However, this method is time consuming, as the network has to be run separately for all windows and much overlap occurs between the patches. In addition, there is a trade-off between the amount of local and global context that is taken into account during classification. [30]
Long et al. [31] proposed to extend the traditional classification networks to the task of segmentation by using Fully Convolutional Networks (FCN). In a FCN, the fully connected layers are removed and replaced by convolutional layers to obtain a likelihood map, which is then upsampled to get per pixel prediction results. [31] This development enabled the network to obtain pixel-wise predictions from entire images instead of patch-wise predictions and made it possible to acquire a prediction for the entire image in just a single pass [32]. This caused the state-of-the-art segmentation performance to improve greatly [33]. Anthimopoulos et al. [34] applied a FCN, called LungNet for semantic segmentation of pathological lung tissue on high-resolution CT-images. Hossain et al. [35] extended LungNet with 3D convolutional blocks to incorporate the 3D information available in CT-scan volumes.

Inspired by the FCN, Ronneberger et al. [30] proposed the U-Net for segmentation, which outperformed the previously mentioned CNN-based model of Ciresan et al. [27] on the same dataset. The U-Net's architecture comprises a downsampling path, which follows the architecture of a CNN, and an upsampling path, consisting of transposed convolutions. The downsampling path, referred to as encoder, aims to capture both spatial and context information. The upsampling path is called the decoder, which is used to regain details and locate the object's position. More importantly, the U-Net contains skip-connections that connect the layers of the decoder with the same sized layers of the encoder. This allows the passage of high-resolution location and context features to the decoder path to regain the spatial information that got lost during down-sampling. [30] A schematic of this architecture is illustrated in Figure 1.1. Generally, segmentation networks, such as the U-Net, are optimized by minimizing the pixel-wise difference between the predicted output and its ground truth. The cross entropy loss is the most frequently used loss function for this application. [36] Other per-pixel loss functions are the L1 loss, which measures the mean absolute error, or the L2 loss, that measures the mean squared error. These errors are measured between each pixel of the segmentation prediction and its ground truth independently and then averaged over all pixels.

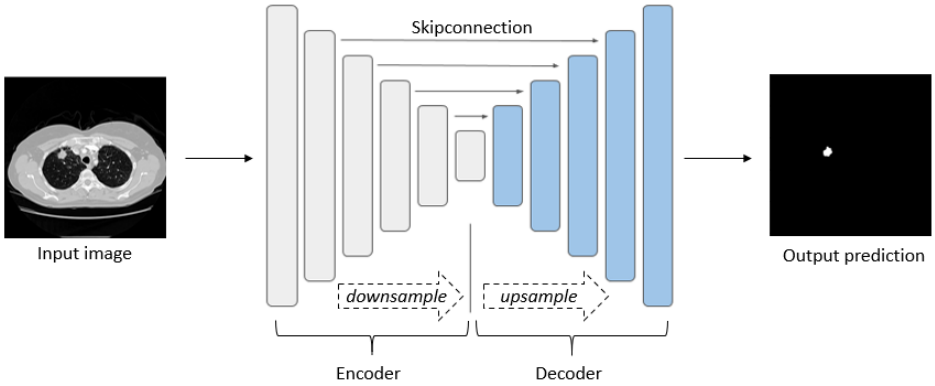


Figure 1.1: Schematic overview of the U-Net architecture, consisting of an encoder (with convolutional and pooling layers) and a decoder (consisting of transposed convolutions), which are connected by skipconnections.

Nowadays, the U-Net is still one of the most used and widely-known structures for medical image segmentation tasks [32], including lung nodule segmentation. Lan et al. [37] suggested a residual U-Net for automatic detection and segmentation of lung nodules, in which the lung parenchyma was segmented prior to nodule segmentation, achieving a Dice overlap score of 71.9%. Keetha et al. [38] developed the U-Det, a modified U-Net that incorporated a bidirectional feature network to segment lung nodules. They obtained a state-of-the-art Dice overlap value of 82.82% on entire thoracic CT-slices. In addition to these papers, many other researchers have used U-Nets as the base of their segmentation methods for patch-based lung nodule segmentation. [39–42]

1.2.3 GANs

Although the aforementioned methods managed to obtain promising results for segmentation, they still suffer from some limitations. Most of the methods are limited by the pixel-wise loss functions, predicting each pixel independently from each other. [43], [33] This causes them to be insufficient in learning both global and local relations between pixels. [44] To alleviate this lack of spatial contiguity (e.g. holes), they require post-processing methods as an additional step, such as connected component analysis or Conditional Random Fields (CRF) [43], [33]. However, this additional post-processing comes at the cost of computational efficiency. [45] Although U-Nets try to overcome this limitation by allowing spatial information passage via skip connections, spatial consistency can-

not be assured in the eventual predicted segmentation. [46] They still rely on a pixel-wise loss and sometimes lack to produce realistic images. [44] For example, blurry results could be produced when minimizing the (Euclidean) distance between ground truth pixels and generated output pixels. [47] These limitations were mitigated by the development of GANs, which made it possible to train a network to generate realistic images in and end-to-end manner, without requiring additional post-processing methods. A GAN consists of two networks; a generator network that produces images and a discriminator network that classifies whether they are real or not. Instead of minimizing a pixel-wise loss, GANs pursue the more general, high-level goal of generating output that is indistinguishable from reality. Blurry images or images that contain discontinuities will not be tolerated since they do not look realistic. [47] In segmentation tasks, the existence of a discriminator is an additional constraint on the generator network, which urges the generator to generate more accurate segmented images. [48]

Since their development in 2014 by Goodfellow et al. [49], GANs have been widely used in the field of biomedical image analysis, such as image synthesis [47, 50], image quality enhancement [51–53], multi-modality image translation (e.g. generating CT-images from MRI images) [54], segmentation [44, 45, 55], classification [56] and detection [57, 58].
Luc et al. [43] were the first to present GANs and adversarial training to the task of semantic segmentation in 2016. They trained a convolutional segmentation network together with an adversarial network with the motivation that it could detect and repair inconsistencies between the predicted segmentations and the ground truths. Isola et al. [47] developed a framework based on conditional GANs for image-to-image translation, which is the widely-known, openly available Pix2Pix framework. This framework can be used for segmentation tasks by translating paired images from one domain (input image) to another (segmented image). GANs have also been studied for segmentation challenges in the medical field. Moeskops et al. [59] used adversarial training to advance CNN-based segmentation of brain MRI scans. They showed that, by including an additional adversarial loss function, inconsistencies could be captured that the normal pixel-wise loss alone did not capture. This resulted in an elevated Dice overlap score compared to training without an adversarial loss function. Kohl et al. [45] proposed to use purely adversarial training by omitting the pixel-wise loss function. They applied a FCN in combination with a discriminator network for semantic segmentation of prostate cancer on MRI images. This purely adversarial approach showed improved segmentation performance than training with a pixel-wise loss function. GANs have also been implemented for pseudo-healthy synthesis, which is the task of generating 'healthy' images from pathological ones. [57, 58, 60] These images can be helpful in tasks like anomaly detection and image segmentation, since additional information is provided about what the image should look like if it were healthy. Moreover, pseudohealthy synthesis is based on unsupervised learning, which eliminates the need for labor-intensive, annotated training data. However, this method is mostly used for detection rather than for precise, pixel-wise segmentation and difficulties arise in images with low contrast differences between healthy and pathological images. [60]

1.2.4 GANs for lung nodule segmentation

For the usage of GANs in segmentation tasks, a discriminator network is added to a segmentation network, e.g. a U-Net. This introduces an adversarial loss based on the real/fake output of the discriminator, which is often combined with a pixel-wise loss function to not only generate realistic segmentations, but also generate segmentations that are near the ground truth. A schematic overview of the general architecture of such a GAN-based segmentation network for lung nodule segmentation is shown in Figure 1.2.
Due to the success of GANs for other medical applications, the implementation of GANs for lung nodule segmentation has recently been explored by few other researchers. Pang et al. [61] developed the CTumorGAN, which is an unified framework based on generative adversarial learning for the segmentation of lung, liver, and kidney tumors on CT-images. They use a U-Net with strided convolutions as generator network and combine three different loss functions in their network: A

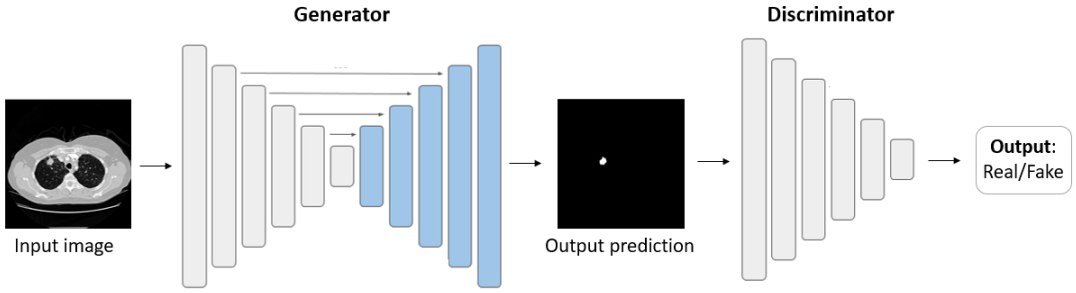


Figure 1.2: Schematic overview of a GAN architecture for segmentation tasks. In this case, the generator network is a U-Net, to which a discriminator network is added.

L2 loss, to push the generator's prediction to correspond to its ground truth; a Dice loss, to overcome the class imbalance; and the adversarial loss, to produce realistic segmentations without fuzzy boundaries. They train and test their network on entire 2D images of the non-small cell lung cancer (NSCLC)-Radiomics dataset to assess the performance of their network on the lung tumor segmentation task. Around the same time, Shi et al. [7] introduced a GAN-based framework for automatic lung nodule segmentation as well. They propose a patch-based network named Aggregation-U-net GAN (AUGAN), in which the generator is a U-Net with deep aggregation (based on U-Net++) and the discriminator is a classification network. Their argument for using a GAN is that adversarial learning could make anatomical tissue details better presented. The loss of their generator includes a L2 loss, a VGG los and the adversarial loss. They demonstrate that adversarial training can be of added value to a U-Net for lung nodule segmentation.

1.2.5 Challenges lung nodule segmentation

Accurate segmentation of lung nodules in CT-images is a difficult task. Three challenges that need to be taken into account, when creating an automatic segmentation method for lung nodule annotation based on GANs, are: (i) The variation in lung nodule characteristics, (ii) the class imbalance that arises with medical image segmentation, and (iii) the instability problems of GANs.

Lung nodule variability There are two main types of lung cancer, non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC), from which the latter is known to grow and metastasize more rapidly. [62] Lung cancer is characterized by the existence of nodules or masses. The definition of a lung nodule is a tissue abnormality that appears on a CT-scan as a roughly rounded opacity with a diameter of up to 3 cm, which can be benign or malignant. [63] Lesions that have a diameter of > 3 cm are known as lung masses and are always considered as malignant, until biopsy proves otherwise. [5]
The detection, classification and segmentation of lung cancer is a complicated procedure due to variability in the appearance of lung nodules on CT-images. They come in various sizes, intensities, shapes, and locations. [6, 50] A major difficulty in lung nodule segmentation is to create a method that is able to take into account these variations in both internal texture as external surrounding. [39] According to variations in internal texture, nodules can be classified into three different types, based on their attenuation in CT-images. (i) Solid nodules, represented by a well-defined shape and a high contrast, (ii) ground glass opacity (GGO), which have fuzzy boundaries and a low contrast to their surrounding, and (iii) part-solid nodules, containing both ground-glass and solid properties. GGO and part-solid nodules occur less frequent than solid nodules, but have a greater possibility to be malignant. [5, 64] Nodules can also be categorised into three groups based on their external surrounding: (i) Well-circumscribed, where the nodule is located completely inside the lung, with no connections to its neighbouring structures, (ii) juxta-vascular, where the nodule is connected to neighbouring vasculature and (iii) juxta-pleural, where the nodule has connections with the pleural

surface. [25] Typical cases for each type of nodule are shown in Figure 1.3.

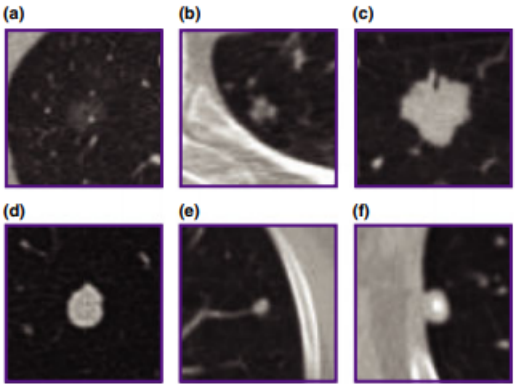


Figure 1.3: Typical examples of each type of nodule. Upper row: Nodules are categorized based on internal texture. (a) Ground glass opacity; (b) part-solid; (c) solid. Lower row: Nodules are categorized based on external surroundings. (d) well-circumscribed; (e) juxta-vascular; (f) juxta-pleural. Source: [39].

The visual similarity of the nodules and their surrounding structures makes it difficult to obtain an accurate segmentation. In case of juxta-pleural and juxta-vascular nodules, the intensity of the nodule is comparable to the intensity of the pleural wall and vasculature, making it difficult to distinguish them from their surroundings. GGO nodules have low contrast to the background, causing conventional methods based on thresholding-and morphological operations to be useless. In addition, small nodules ($<$4 mm) are hard to distinguish due to their size and intensity, which is comparable to surrounding noise. [38] Moreover, there is much variety in lung nodule size, with diameters ranging from 3 to 30 mm [39] Therefore, conventional methods based on image processing techniques, which are often solely useful for the segmentation of well-circumscribed solid nodules, cannot be simply applied to all types of nodules. [29]

Class imbalance Class imbalance is often a severe problem in medical image segmentation and is caused by an uneven distribution of the classes in the data. This occurs if pixels corresponding to a common class are far more numerous than pixels corresponding to a minority class. This causes the learned model to classify most of the pixels as members of the majority class, as the rarity of the minority class inhibits accurate learning and labeling. [65] The segmentation of lung nodules in CT-images is one of the applications that is affected by this class imbalance, as the volume of a lung tumor is often significantly smaller than the volume of healthy tissue. This causes pixels in the tumor class to be outnumbered by the pixels in the background class. Consequently, the tumor region is often missing or partly detected. Shi et al. [7] addresses this problem by training their GAN based network on patches of lung nodules. This reduces the amount of background pixels and therefore decreases the class imbalance. However, it is desirable to create a network that takes entire images as input, to avoid the need of detecting the lung nodules prior to segmentation. Pang et al. [61] used an additional Dice loss in their CTumorGAN to overcome this problem. This loss was first introduced by Milletari et al. [66] and was adapted from the Dice coefficient, one of the most prevailing measures for region overlap in image analysis. [66] It was advantageous over previous approaches that used loss functions based on sample re-weighting [30, 31]. In these approaches, foreground regions were allocated more importance than background pixels, but the choice of weights was hard to optimize and inappropriate selection could easily cause bias towards rare classes. [66]

GAN training instability Despite their successes, GANs face many problems, of which training instability is the most substantial. [67] The emergence of these problems will be discussed in more

detail in section 2.2. Various methods have been suggested to address the instability problems during training. However, all these methods rely on the single value yielded by the discriminator, which categorizes the entire image as real or fake. Since medical image segmentation involves dense, pixel-level labeling, only the real/fake classification could be too easy for the discriminator. [44] This results in an unstable network with lack of sufficient gradient feedback to improve the generator network. Salimans et al. [68] proposed an adversarial feature matching loss for the generator to stabilize adversarial training. Instead of only maximizing the output of the discriminator, this loss ensured that the features, extracted at each layer of the discriminator network, matched between the real data and the generated data. Xue et al. [44] adopted this feature matching loss and extended it by using this loss for the generator as well as the discriminator (i.e. critic) network. They proposed an architecture called SegAN, which employs an adversarial multiscale L1 feature loss that captures the difference between the generated segmentation and its ground truth at each layer in the critic network. This forced both the segmentation and critic network to learn discriminative global and local features. The multiscale L1 feature loss prevents the gradient feedback to the networks from vanishing, making the adversarial training stable, as well as avoiding the segmentation network from overfitting. [69] Since its first application by Xue et al. a number of researchers have employed the multiscale L1 loss for segmentation tasks. [70, 71]

1.3 Objective

The goal of this thesis is to implement a GAN-based segmentation method that is capable of annotating lung nodules and to investigate whether this method could be of added value for improved segmentation performance compared to the existing techniques. The main research question this work is trying to answer is:

To what extent can a GAN-based network be used for automatic segmentation of lung nodules in entire 2D CT-images?

The method proposed in this thesis has the intention to overcome the three aforementioned segmentation challenges. It is a GAN-based network with a U-Net as generator network, to which three main modules are added:

- To address the problem of nodule size variations and their correspondence to the surroundings, parallel dilated convolution modules are added to capture both local and global context.
- In order to solve the problem of class imbalance, an additional Dice loss is employed, which focuses on maximizing the overlap between the ground truth and the predicted segmentation.
- The classic discriminator network of a GAN is replaced by a critic network that extracts a multiscale L1 loss at all the layers of the critic's network. Using a multiscale loss function could prevent instable and insufficient gradient feedback to the network, which can make the network train more stably than normal GANs.

The sub-questions that will help answer the main research question are:

1. To what extent do the additional modules solve the segmentation challenges and contribute to a more accurate segmentation performance of the proposed method?
2. How does the performance of the proposed method compare to the other GAN-based methods for lung nodule segmentation?
3. What are the performance differences between the proposed method and the U-Net and could the proposed method be of added value for the task of automatic lung nodule segmentation compared to this state-of-the-art segmentation network?

1.4 Report outline

The remainder of this report is structured as follows. Chapter 2 will provide some background information on GANs. It will explain the working principle of a GAN, the challenges GANs face, and the various types of GANs. The method used in this thesis will be described in Chapter 3, covering a description of the database used, the proposed architecture, and the experimental set-up. The results of the experiments will be demonstrated in Chapter 4. These results will be discussed and related to other studies in Chapter 5. Finally, Chapter 6 will conclude this thesis and some recommendations will be given towards future work in this area of research.

2 Theoretical background GANs

In this study, GANs are used as an extension of the U-Net by the addition of a discriminator network. To understand how GANs can be used for image segmentation tasks, we first need to understand the basics of the conventional GAN. The relevant theoretical background is provided in the current chapter. Section 2.1 covers the theory and working principle of GANs. The main limitations the original GANs face and possible solutions are described in Section 2.2. Lastly, various types of GANs that could be of interest for this study are given in Section 2.3.

2.1 Working principle

GANs consist of two competing networks. The generator network focuses on the generation of new data by learning the underlying distribution of a given dataset, often images. This network is optimized with a second network, known as the discriminator, which aims to distinguish whether the provided data is sampled from the generated data distribution or the real data distribution. The result is adversarial training: the discriminator's objective is to discern the real and the fake images, while the generator aims to produce output images that can confuse the discriminator. [13]

In Figure 2.1, a schematic overview of a GAN network is shown. The generator network (G) is a differentiable function controlled by some set of parameters θ_g. G transforms noise vectors z from a random distribution p_z into a distribution p_g with samples $\hat{x} = G(z)$. The generator's goal is to generate perceptually convincing samples that resemble the real data distribution p_{data}. In other words, it tries to minimize the difference between p_{data} and the generated distribution p_g. The discriminator network is introduced to be able to determine this difference. The discriminator (D) is a differentiable function controlled by parameters θ_D, and can be considered as a binary classification network. D is provided with real samples x from p_{data} and generated samples $\hat{x}$, and tries to differentiate which ones originate from the real data distribution and which ones originate from the synthesized data distribution. [49]

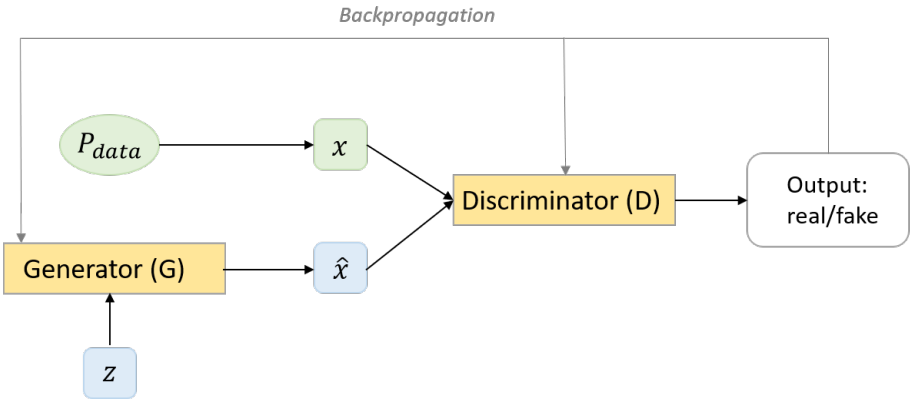


Figure 2.1: Schematic representation of a GAN. The generator takes a noise vector z as input, which was sampled from a random distribution p_z, and transforms it into an output sample $\hat{x}$. The discriminator attempts to categorize these samples as fake and the samples drawn from the real data distribution, p_{data}, as real. Both the discriminator and generator are optimized through backpropagation of the discriminator's output

2.1.1 Training process

GAN training is performed by sequentially updating the discriminator and the generator. Both the discriminator and the generator try to minimize their own cost function; $J^D(\theta_D, \theta_G)$ and $J^G(\theta_D, \theta_G)$ respectively. These cost functions are defined by both the networks' parameters, θ_G and θ_D, but

the networks are only able to adjust their own parameters. In the training procedure, two phases can be identified; updating the discriminator D and updating the generator G. In the first phase, D is updated by introducing real samples drawn from p_{data} to D. The output, $D(x) \in [0, 1]$, is the probability that a real sample x came from p_{data}. Thereafter, D is updated using fake samples drawn from p_g. The output, $D(G(z))$, is a probability in the range of [0,1] that a fake sample $\hat{x}$ originated from p_{data}. The cost function measures the discrepancy between the predictions and their target labels (0 for fake samples, 1 for real samples). The error is then backpropagated and the parameters of the discriminator network are updated accordingly. In this phase, the parameters of the generator are kept fixed. [49]
Secondly, G is updated with a minibatch of noise vectors drawn from p_z. The output of D is the probability that $G(z)$ (i.e. $\hat{x}$) originated from p_{data}. G tries to minimize the probability that a sample $\hat{x}$ is classified as fake. The loss is backpropagated to G to update the parameters of the generator network, while the discriminator is fixed. [49]
The parameters of the networks are updated using an iterative stochastic gradient descent (SGD) optimization method with two gradient steps: one updating θ_D to reduce J^D and one updating θ_G to reduce J^G. Different gradient-based optimization algorithms could be used, of which the Adaptive Moment Estimation (Adam) optimizer is usually a good choice. [67]

2.1.2 Objective function

The GAN training situation, in which the cost functions are defined by both the networks' parameters but the networks are only able to adjust their own parameters, is often described as a non-cooperative, zero-sum game to which a Nash equilibrium is the solution. In this context, a Nash equilibrium is reached when a point (θ_G, θ_D) has been found such that both J^G and J^D reside in a local minimum. [67] This results in a two-player minimax game, where the discriminator tries to maximize the following objective function and the generator tries to minimize it [49]:

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

In Equation 2.1, $V(D, G)$ is the objective function, p_{data} is the real data distribution and p_z is the random distribution from which the noise vectors are extracted. This objective function can be derived from the binary-cross entropy (BCE) loss, which is defined as:

$$L(\hat{y}, y) = -[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})] \tag{2.2}$$

Where $\hat{y}$ is an output prediction and y is the target label.

In case of training D with real samples, the label of the data coming from $p_{data}(x)$ is $y = 1$ and the predicted output $\hat{y} = D(x)$. By substituting this in the equation of the BCE loss function, the resulting loss function becomes:

$$L(D(x), 1) = -\log(D(x)) \tag{2.3}$$

When updating the discriminator with samples from the generated data distribution p_g, the input label is $y = 0$ and the predicted output $\hat{y} = D(G(z))$. This results in the loss function:

$$L(D(G(z)), 0) = -\log(1 - D(G(z))) \tag{2.4}$$

The loss functions above are only valid for one single sample. When considering the entire dataset, the functions must be converted into expectations. Taking this into consideration and combing the equations above, the loss function of the discriminator is given as:

$$J^D(\theta_D, \theta_G) = -[\mathbb{E}_{x \sim p_{data}} \log(D(x)) + \mathbb{E}_{z \sim p_z} \log(1 - D(G(z)))] \tag{2.5}$$

The objective of the discriminator is to correctly categorize the data from $p_{data}(x)$ as real and the data from $p_g(z)$ as fake. Instead of minimizing the the loss, the objective of D can also be achieved by maximising the opposite of the loss, resulting in:

$$\max_{\theta_D}[\mathbb{E}_{x\sim p_{data}}\log(D(x)) + \mathbb{E}_{z\sim p_z}\log(1 - D(G(z)))] \tag{2.6}$$

After updating the discriminator for one minibatch, the generator is updated. The generator competes with the discriminator, attempting to minimize the probability that the discriminator will classify a sample drawn from $p_g(z)$ as fake, i.e. $y = 0$. Therefore the generator's loss relies directly on the discriminator's performance:

$$J^G(\theta_D, \theta_G) = \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \tag{2.7}$$

The generator's objective is to maximize $D(G(z))$ and thus to minimize the loss function:

$$\min_{\theta_G}[\mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]] \tag{2.8}$$

By combining the objective functions for the generator and the discriminator, the overall loss function can be written as Equation 2.1.

The iterative training process of GANs tends to be unstable, which makes the optimization of the networks challenging. In practice, Equation 2.8 may not provide sufficient gradients for the generator to train well. [49] In the early stages of the process, the discriminator can easily differentiate the real and fake samples. In that case, the gradient for the generators loss function is close to zero. As a result, the generator will not be able to adjust its parameters and minimize the loss. Generally, this is solved by applying an alternative loss function for the generator:

$$\min_{\theta_G}[-\mathbb{E}_{z\sim p_z}[\log(D(G(z)))]] \tag{2.9}$$

In this equation, the goal of the generator is to maximize the probability that the discriminator classifies the generated samples as real, instead of minimizing the probability of classifying the generated samples as fake. This objective function especially provides much stronger gradients in early training. Therefore, this type of GAN is called the Non-Saturating GAN (NS-GAN). [49]

2.1.3 Theoretical results

The minimax game of Equation 2.1 has a global optimum for $p_g = p_{data}$ [49], which can be proven by the derivation of the optimal discriminator and optimal generator. This section shows a concise version of the derivation. The reader is referred to Appendix A for the extensive derivation of the optimal discriminator and generator and the theoretical solution.

The optimal discriminator D^* for fixed G is:

$$D^*_G(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \tag{2.10}$$

The optimal G, G^*, can be calculated by minimizing Equation 2.1 with respect to G and implementing the optimal discriminator:

$$G^* = \min_G V(D^*_G, G) \tag{2.11}$$

This problem can be solved using the Jensen-Shannon-Divergence (JSD) [72], resulting in:

$$V(D^*_G, G) = -\log 4 + 2JSD(p_{data} \parallel p_g) \tag{2.12}$$

The JSD between two distributions is always positive and zero only when the distributions are equal. [72] Therefore, the global minimum of equation 2.12, and thus the solution for G^*, is $-\log 4$. This is achieved if and only if $p_g = p_{data}$. The optimal values for D and V then are:

$$\begin{aligned} D^*_G(x) &= \frac{p_{data}}{p_{data} + p_g} = \frac{1}{2} \\ \min_G \max_D V(D,G) &= \mathbb{E}_{x\sim p_{data}}[\log \frac{1}{2}] + \mathbb{E}_{z\sim p_z}[\log\,(1 - \frac{1}{2})] \\ &= -2\log 2 \end{aligned}$$

At this point, both D and G have reached a situation in which they cannot improve. The discriminator then is unable to differentiate between the two distributions. [49]

2.2 Challenges GAN training

GANs are a powerful technique for image generation in many fields. Nonetheless, the original GAN suffers from instability problems due to non-convergence, such as vanishing or exploding gradients and mode collapse. [67, 73] As a consequence, the process of training GANs is a challenging task. Finding balance between the generator and discriminator is of great importance for stable training. However, because both networks are simultaneously trying to minimize their own costs, there is no guarantee of reaching a Nash equilibrium, provoking a chance that one network becomes more powerful than the other. [67] Most often the discriminator becomes too strong compared to the generator. The generated segmentations then become easily distinguishable from real ones, causing the gradients from the discriminator to approach zero. Without gradients, there is no guidance for further training the generator. [46]
Another problem is mode collapse, which occurs when the generated distribution $p_g(x)$ learned by the generator concentrates only on a few modes of the distribution $p_{data}(x)$. Consequently, the generator produces only a small subset of samples, instead of generating diverse images, resulting in poor output diversity. [49] Complete mode collapse is rare, but partial mode collapse is very common, e.g. when the generator creates multiple images, but they all contain the same texture or color. Mode collapse may arise because simultaneous gradient descent does not prioritize minmax over maxmin. [67] Although this seems like a minor difference, the maximin solution to the GAN game differs considerably from the minimax solution. When the model behaves like a maximin game instead of a minimax game:

$$G^* = \max_D \min_G V(G,D) \tag{2.13}$$

the generator will learn to map every noise vector z to the one x sample that the discriminator most likely considers as real rather than fake, instead of mapping every z vector to different x samples. Therefore, it tries to optimize and perfect one mode, instead of performing well, but average, on all modes. [67]

2.2.1 Solutions

Various solutions have been proposed that are motivated to encourage convergence and thus solve training instability. Three main solutions are the use of mini-batch features, applying one-sided label smoothing and using feature matching. [68] Beside these solutions, adapting the objective function could also improve training stability. This is further explained in Section 2.3.2.

Mini-batch features An attempt to mitigate the mode collapse problem is to use mini-batch features, which allows the discriminator to compare each sample to a minibatch of generated samples and a minibatch of real samples. This provides the discriminator the ability to detect if the sample is unusually similar to other generated samples. If so, the discriminator can penalize the generator to avoid collapse of the generator. [68]

One-sided label smoothing One-sided label smoothing is a regularization method to prevent overconfident predictions of a neural network. [74] The discriminator of a GAN is prone to producing extremely confident outputs that identify the correct class, but with a too extreme probability, which is not optimal for the generator. By replacing the label for real images with a value slightly less than 1, e.g. 0.9., the discriminator is encouraged to estimate soft probabilities instead of extremely confident classification. The discriminator is then penalized when its prediction for any real image goes beyond 0.9. This technique prevents the discriminator from returning extremely large gradient signals to the generator. It is important to mention that only the labels for real data are smoothed, not for the fake data, as this could reinforce incorrect behaviour in the generator. When the label for fake data is non-zero, the optimal discriminator function (Equation 2.10) changes: p_g enters in the numerator of the function. The presence of p_g in the numerator is problematic because, in regions where p_{data} is very small and p_g is larger, fake samples from p_g will not be stimulated to move nearer to the real data distribution. Hence, only the positive labels are smoothed. [67]

Feature matching Feature matching is a method that addresses GAN instability by introducing a regularization objective for the generator. The new objective function is defined as: $||\mathbb{E}_{x \sim p_{data}} f(x) - \mathbb{E}_{z \sim p_z} f(G(z))||$, where f(x) represents any statistics computation of intermediate layer features, such as mean or median. This objective function pushes the generator to produce data that corresponds to the statistics of the real data. The generator network is trained to match the statistics of real data features on any layer of the discriminator. The discriminator is used to extract these feature vectors at each intermediate layer. By training the discriminator, it is asked to find those features that are especially discriminative of real data versus generated data. This ensures the generated images to have features that resemble the real images. It has been shown that feature matching is effective in situations where an original GAN becomes unstable. [68]

Beside the solutions mentioned above, numerous improvements to the original GAN have been proposed, both in architecture as in loss functions, which will be further elaborated in the next section.

2.3 Types of GANs

As GAN research has obtained increased interest over the last years, numerous variations on the original GAN's architecture and loss function have emerged. Mentioning all the variations is beyond the scope of this study, so only the ones that could be of interest for this study will be pointed out in the next sections.

2.3.1 Variations in architecture

Initial applications of GANs adopted multilayer perceptrons (MLPs) for the generator and discriminator network. Although these MLPs could be sufficient for smaller images, these networks weren't suitable for the synthesis of larger images, such as medical images. [13] Therefore, these multilayer perceptrons were replaced by convolutional neural networks. Nowadays, multiple architectural variations on the original GAN exist. For this study, two architectural variations are especially of interest: the Deep Convolutional GAN and the conditional GAN.

Deep Convolutional GAN Deep Convolutional GANs (DCGAN) were introduced by Radford et al. [75] in 2015 and most of the GANs nowadays are (loosely) based on the DCGAN architecture. [67] Although GANs were already deep and convolutional prior to the introduction of DCGANs, the name refers to a specific style of architecture. The architectural restraints ensured stable training and allowed for training deeper and higher resolution generative models. [75] The architectural adaptions for stable DCGANs were:

- Pooling operations were replaced with strided (transposed) convolutions in both D and G.

- Batch normalization was applied to most layers in D and G.
- A Rectified Linear Unit (ReLU) activation was used in G. Only the output layer used a hyperbolic tangent (Tanh) activation.
- All fully connected hidden layers were removed for deeper architectures.
- In D, LeakyRelu was used as activation in all layers.
- The Adam optimizer is used instead of SGD.

Conditional GAN For segmentation applications, it is necessary to be able to control the generated output, as the segmentation must correspond to the input image. The original GAN was not suited for this task, and therefore conditional GANs (cGANs) were introduced to segmentation challenges. [43] CGANs are widely applied for various image generation tasks, among which segmentation is one. They were first proposed by Mirza et al. [76] and allow, in contrast to the original GAN, control over the characteristics of generated samples by adding an additional condition (c) to the generator and discriminator network. This is shown in the schematic of Figure 2.2.

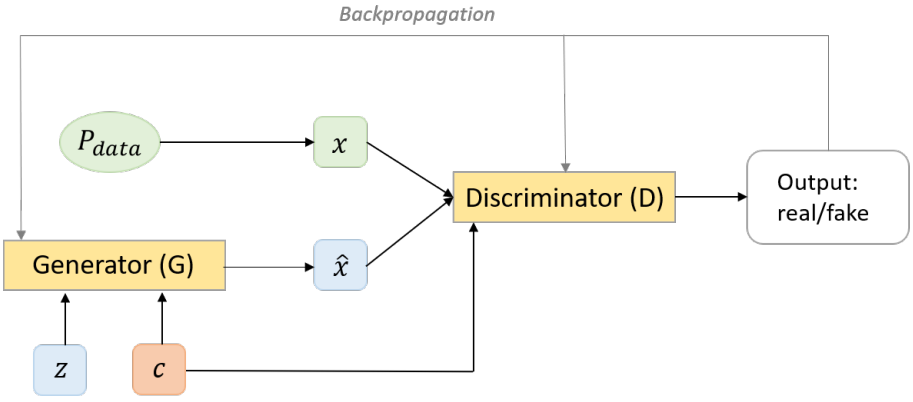


Figure 2.2: Schematic representation of a conditional GAN. In addition to the noise vector z, the generator takes an additional condition c as input, which is also inserted into the discriminator

If paired images are available, which is the case for segmentation tasks where input images and annotated ground truths are available, cGANs can be used to translate images from one domain to another domain, i.e. from input image to segmented image. In this case, noise vector z can be omitted and the condition c is equal to the input image (y), making it an image-to-image translation task. [47] The objective function of Equation 2.1 then changes to:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}, y\sim p_y}[\log D(x|y)] + \mathbb{E}_{y\sim p_y}[\log (1 - D(G(y)|y))] \tag{2.14}$$

Where y is the input image and x the corresponding ground truth segmentation. Aditionally, it is regarded beneficial to add a more traditional, pixelwise loss (e.g. L1 or L2 loss) to this GAN objective, to not only generate perceptually convincing images, but also generate images that are near the ground truth output. [47] A widely-known framework for image-to-image translation tasks based on cGANs is the Pix2Pix model. [47]

2.3.2 Variations in loss function

Unstable GAN training is partly caused by the way in which the discriminator measures the difference between p_{real} and p_g, namely by computing the Jensen-Shannon (JS) divergence to express the distance between these probability densities. [13] The JS divergence is a smooth, symmetric divergence measure, but does not provide usable gradients when two distributions are non-overlapping. Therefore, the JS divergence is not applicable for computing the distance between distributions

that have disjoint parts. [77] Alternative objective functions have been proposed to address this limitation by using a different divergence, such as f-divergence (f-GAN) [78], Pearson χ^2 divergence [77] and Wasserstein distance [73], among which Wasserstein GAN (WGAN) is the most popular approach. [46]

Wasserstein GAN Arjovski et al. [73] proposed to use a new distance measurement method in GANs to determine the divergence between the real and fake distribution: the Earth Moving Distance or Wasserstein distance. It can be interpreted as the minimum energy cost to convert one data distribution into another. A WGAN objective function is defined as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}}[D(x)] - \mathbb{E}_{z\sim p_z}[D(G(z))] \tag{2.15}$$

Discriminators in WGANs are often referred to as critics, as they do not return a real/fake probability, but a scalar value that quantifies the similarity between two distributions. [13] Compared to the original GAN, the WGAN has two more adaptions: the Adam optimizer is replaced by a RMSProp optimizer and weight clipping or, more commonly, gradient penalty is used. [73] The advantage of Wasserstein distance compared JS divergence is that when the distributions are non-overlapping, the Wasserstein distance is still able to reflect their distance and vanishing gradients are avoided. [73]

Least Squares GAN Least Squares GAN (LSGAN), proposed by Mao et al. [77], adopts a least squares loss for both the generator and the discriminator. They can provide stronger gradients to update the generator since they penalize samples by their distance to the decision boundary. Therefore, disjoint samples can be pulled closer to the boundary, even when correctly classified. This mitigates the vanishing gradient problem and results in more stable training. By minimizing the objective function of a LSGAN, the Pearson χ^2 divergence is minimized, which equates a L2 loss. [77] The objective functions for LSGANs are formulated as:

$$\min_D V_{LSGAN}(D) = \frac{1}{2}\mathbb{E}_{x\sim p_{data}}[D(x)-b)^2] \; + \frac{1}{2}\mathbb{E}_{z\sim p_z}[D(G(z))-a)^2] \tag{2.16}$$

$$\min_G V_{LSGAN}(G) = \frac{1}{2}\mathbb{E}_{z\sim p_z}[D(G(z))-b)^2] \tag{2.17}$$

Where a and b represent the labels for fake samples (0) and real samples (1), respectively.

3 Method

This chapter covers the details about the database and preprocessing strategy (Section 3.1), the proposed network architecture (Section 3.2), the training procedure (Section 3.3), the experimental set-up (Section 3.4) and the evaluation for performance assessment of the networks (Section 3.5).

3.1 Database

In this study, the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) public database was used for training and validation of the proposed segmentation methods. This openly available database was initiated by the National Cancer Institue and is intended to facilitate the development, training and evaluation of CAD-methods for lung cancer diagnosis and detection. The LIDC-IDRI database consists of 1018 thoracic computed tomography scans with corresponding annotated nodules, which were collected retrospectively from seven academic centers. The nodules were manually annotated by four experienced radiologists in a two-stage process. In the first, blinded-reading phase, the radiologists independently reviewed each CT-scan and annotated suspicious lesions. Each lesion was categorized as nodule $\geq 3\ mm$, nodule $< 3\ mm$ or non-nodule. Manual segmentation was only performed for the lesions classified as nodule $\geq 3\ mm$. In addition, the radiologists also rated the nodules on a scale of 1-5 for nine common characteristics: calcification, sphericity, texture, internal structure, subtlety, lobulation, margin, spiculation and malignancy. In the second, unblinded-reading phase, the annotations of the other three radiologists were provided and each radiologist could decide to modify their original annotations based on the annotations of their colleagues. The aim of this two-stage reading process was to improve the quality of the ground-truth annotations and to identify all the nodules in the scans as completely as possible without needing to force consensus among the radiologists. [79]

3.1.1 Data preprocessing

Multiple preprocessing methods needed to be applied before the images could be accepted for the task of segmenting lung nodules. For retrieving and analysing the data, Python's pylidc library was used, which is an object-relational mapping for the data provided in the LIDC dataset. [80] In the database, there were numerous redundant slices that could be discarded. Moreover, the images in the database were collected by different imaging devices from multiple institutions. Hence, it was essential to implement a normalization procedure. The necessary preprocessing steps applied in this study are described more extensively in the paragraphs 'Slice and nodule selection', 'Ground truth masks preparation' and 'Normalization'.

1. Slice and nodule selection

The 1018 CT-scans available in the LIDC-IDRI database all comprise multiple slices. The slice thickness of the scans varies from 0.6 mm to 5.0 mm. In consideration of image quality, scans with a slice thickness greater than 2.5 mm were excluded since these were not recommended for CAD analysis. [39, 81] In total, 897 CT-scans were included. Each scan consists of a series of 2D image slices with a size of 512 x 512 pixels. A majority of the slices could be discarded as they did not include any nodule. For our application, only slices that contained nodules categorized as $\geq 3\ mm$ were taken into consideration. Nodules smaller than 3 mm were not included since they were not marked by a contour, but by their central point, which is not sufficient for the formation of a ground truth segmentation. In addition, only the nodules accepted by a majority of the radiologists, i.e. at least 3 out of 4 radiologists, were selected as valid nodules. This resulted in a set of 1179 nodules. Since a nodule is a 3D structure, it spans multiple slices. Instead of only including the center slice of a nodule, as many applicable slices of one nodule were included to obtain the largest possible dataset. Only the slices in which the nodules contour was (1) marked by at least 3 radiologists and (2) still reached a diameter of $\geq 3\ mm$, were included. This amounted to a final dataset of 6915

images emanating from 1179 nodules. In Figure 3.1, the histogram of nodule amount and diameter distribution is shown.

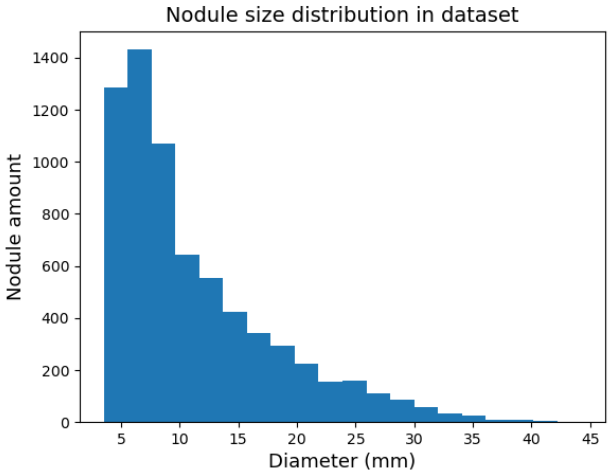


Figure 3.1: Histogram of lung nodule diameters across the processed dataset.

2. Ground truth masks preparation

Once the correct slices were selected, the ground truth (GT) masks were created by combining the annotations from all radiologists into a single GT boundary. This was based on a 50% consensus criterion, which implies that two or more radiologists should have included a given pixel in the nodule boundary to mark that pixel as part of the consensus contour. An example of a nodule annotated by all four radiologists and its corresponding 50% consensus boundary is shown in Figure 3.2. All pixels inside the 50% consensus boundary were set to 1 and all pixels outside this boundary were set to 0 to create the binary GT masks.

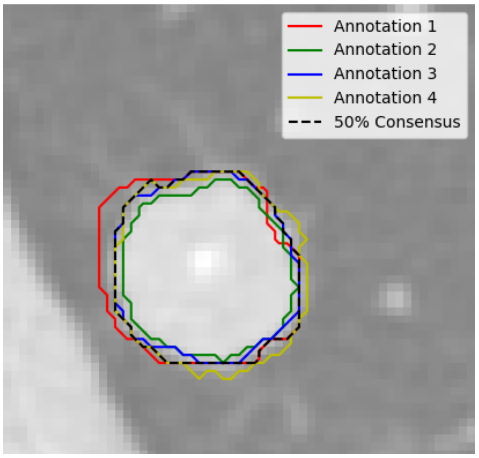


Figure 3.2: Lung nodule annotated by all four radiologists and its corresponding 50% consensus boundary.

3. Normalization

The images in the database were collected by different imaging devices and different acquisition protocols, causing a variety in pixel spacing and intensities to occur. A normalization procedure was applied to achieve equal pixel spacing and intensity values in all images. To maintain as much original information as possible, the resampling step preceded the intensity normalization step in order to prevent interpolation with normalized intensity values.

- **Pixel spacing** The pixel spacing in the entire dataset varied from 0.461 to 0.977 mm, with an average spacing of 0.688 mm. The images were resampled to achieve the same, average pixel spacing of 0.688 mm. Resampling was applied to the input images and the GT masks using second order b-spline interpolation and nearest-neighbour interpolation, respectively. The difference in interpolation method between the input image and GT image arises due to the difference between continuous and binary images. For continuous images, B-spline interpolation appeared to be optimal, while for the discrete, binary GT images, the usage of nearest-neighbour interpolation was crucial to obtain accurate results. Equation 3.1 gives the definition of pixel spacing, in which the physical size equals the real size of the imaged object. From this formula it can be seen that, due to the resampling procedures, images with an original pixel spacing smaller than the average pixel spacing will become smaller than 512×512 pixels, whereas images with a pixel spacing larger than the average pixel spacing will become larger than 512×512 pixels.

$$\text{pixel spacing} = \frac{\text{physical size}}{\text{number of pixels}} \tag{3.1}$$

  In order to obtain same-sized images, the images smaller than the original size were enlarged by applying zero padding until they re-reached the original size of 512×512 pixels. The images that became larger were cropped to this size. A size of 512×512 was chosen as this was the original size of the images. We also experimented with smaller sizes, since this would be beneficial to the computation time, but this caused relevant structures of the lungs to be lost during cropping.

- **Pixel intensity** For the same reason as for the pixel spacing, the pixel intensity was normalized as well. Pixel intensities in CT-images are given in Hounsfield Units (HU), which depicts the relative density of organs and tissues on CT-images. [82] The HU values varied greatly between the different CT-images. After analysis, the meaningful pixel intensities for nodule segmentation were determined to lie in the range of [-1000, 400 HU]. It was essential to use appropriate window settings to avoid missing lesions and structures of interest. Therefore, the range was chosen based on experimental analysis as well as on literature findings. Mukhopadhyay et al. [83] noted that the HU-values of non-solid nodules are ranged between -750 and 300 HU, whereas solid nodules have a density distribution range of [-200, 200 HU]. Part solid nodules are a mixture of solid and non-solid nodules and thus fall in between these ranges. [83] Moreover, for lung images, a window setting of level -600 HU and width 1000 HU is often used to appropriately visualize the lungs. [84] Based on these results, the range of [-1000, 400 HU] was considered appropriate for our application. For each CT-slice, the intensities were first clipped in the selected range. Then, the pixel values were normalized to [-1,1].

3.1.2 Characteristics

The dataset was randomly partitioned into three subsets for training, testing, and validation with a partition rate of 0.8, 0.1 and 0.1 respectively. This corresponds to 5533 images in the training dataset and 691 images in both the validation and test set. The distribution of the nodules' subtlety scores in each subset is plotted in Figure 3.3. The subtlety score represents the difficulty of detection and ranges from 1-5. Higher values indicate easier detection. Nodules that are difficult to detect, which are expected to be the small nodules, GGO nodules and adhesion-type nodules (juxta-vascular and juxta-pleural nodules), are graded with a low subtlety score. The nodules were graded by multiple radiologists, causing the final subtlety score to be the mean score of all radiologists. As can be seen in Figure 3.3, the subtlety scores are equally distributed between the three different datasets, which is of importance for accurate training and testing of the proposed method. However, the distribution of the subtlety scores in each dataset is imbalanced; higher scores are more frequently occurring than lower scores. This is a common problem in medical image databases [85], which

could be solved by generating additional images of the underrepresented classes, but this is beyond the scope of this study.

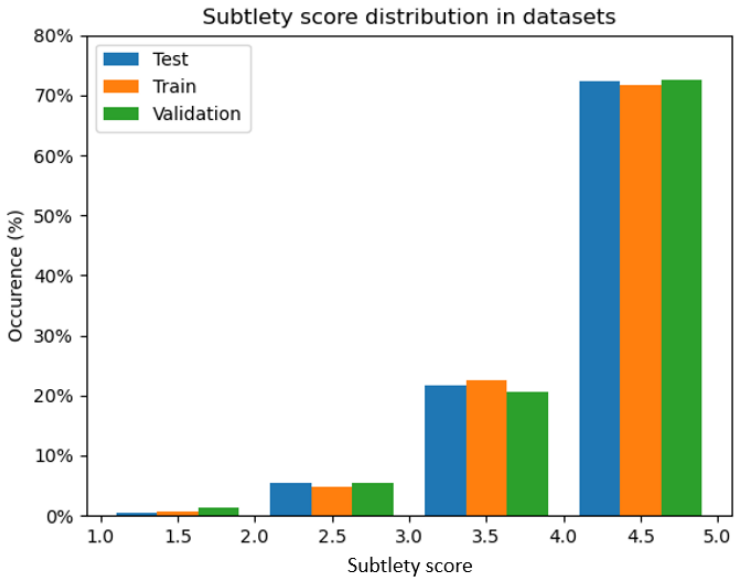


Figure 3.3: Subtlety score distribution in train, test and validation dataset

3.1.3 Corrected dataset

At the end of this study, it appeared that the dataset described above was not completely correct, as the GT masks were created for each nodule separately. This led to erroneous results for slices in which multiple nodules occurred: The network would segment nodules correctly, but due to the GT only including one of the nodules, it would obtain a falsely low performance score on such slices. Therefore, a corrected dataset was created in which the GTs of nodules that appeared in the same slice were superimposed. This resulted in a dataset of 6663 images, which was split into a training, test and validation set of 5331, 666 and 666 images, respectively.

3.2 Proposed architecture

The method proposed in this thesis aims to overcome the three segmentation problems that arise in the task of lung nodule segmentation, which were described in Section 1.2.5. The overall network follows the structure of a conditional image-to-image translation GAN, which is conditioned on the input CT-image. The method is an end-to-end architecture trained on entire 2D input images instead of patches, averting the need to detect and extract lung nodules first. Therefore, it combines the detection and segmentation procedure in one framework. The proposed network comprises multiple subcomponents: a generator network, a critic network, an additional loss function and atrous spatial pyramid pooling (ASPP) modules. The overall proposed architecture is shown in Figure 3.4. The three modules proposed in Section 1.3 to alleviate the segmentation challenges (multiscale L1 loss, Dice loss, and parallel dilated convolutions (ASPP)) are shown as orange boxes. All subcomponents will be further explained in the next sections.

3.2.1 Generator network

The generator network is based on the U-Net of Ronneberger et al. [30] The skip connections in the U-Net allow fusion of low-level spatial information with the high-level features to increase the tumor localization accuracy. [30] Contrary to the basic U-Net, the generator network in the proposed method exploited strided convolutions instead of pooling operations to reduce the information loss

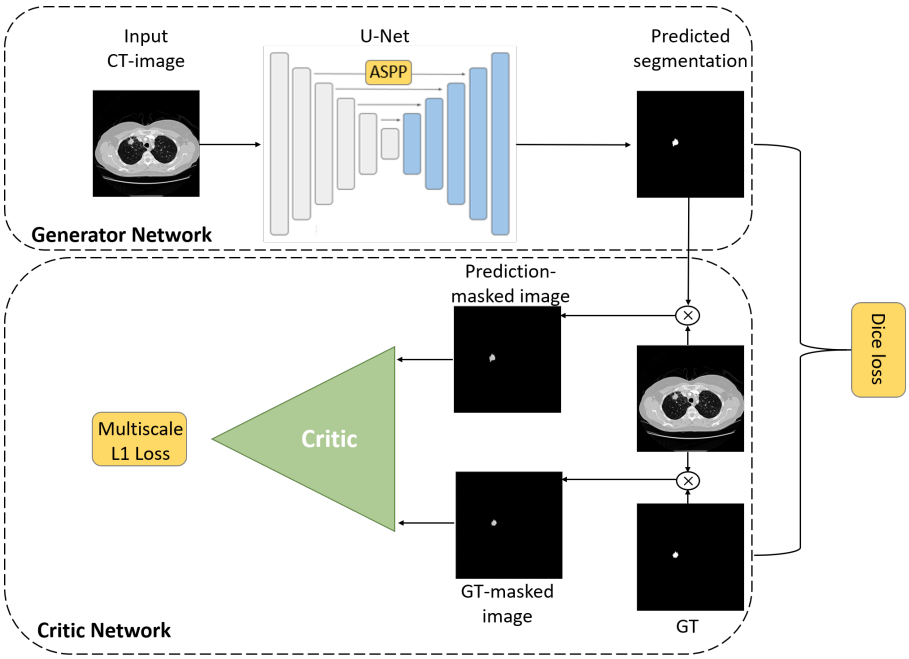


Figure 3.4: Schematic representation of the proposed GAN architecture. The three proposed modules are shown in light orange boxes; Multiscale L1 loss, Dice loss, ASPP module. GT = Ground Truth

from pooling operations, which are sensitive for small tumors. [61] Downsampling was realized by convolutions with stride two and upsampling by transposed convolutions with stride two. Moreover, the generator was enhanced with residual units in each layer of the encoder path and upsample units in the decoder path, which were inspired by the residual units of ResNet that made it possible to train deeper networks. [86] Figure 3.5 shows the residual and upsample units.

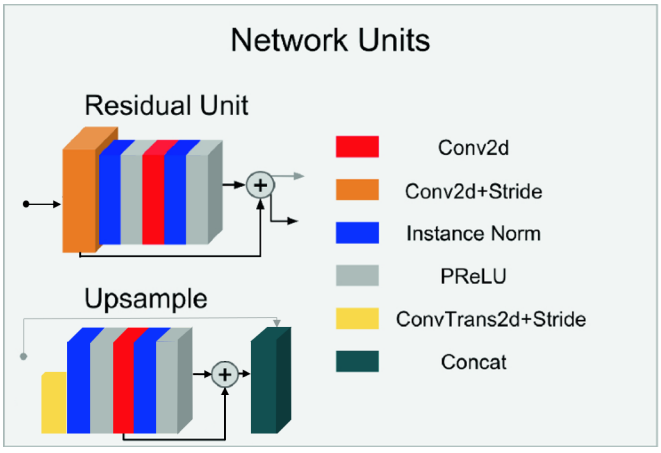


Figure 3.5: Architecture of the residual unit and the upsample unit. Image adapted from [87]

Parametric rectifying linear units (PReLU) were implemented as activation function, which are mathematical functions that determine the output of a neural network. These functions are attached to each neuron in a network, determining whether the neuron should be activated or not, based on the relevance of the neuron's input for the output prediction. PReLU activation counteracts the dying gradient problem of ReLU activations as it does not have zero-slope parts and generalizes it by making the coefficient of leakage a learnable parameter. [88] Therefore they allow the network to learn a more effective activation. [87] As a normalization technique, instance normalization was applied. Normalization is used to standardize the input of each intermediate layer, by fixing the mean and standard deviation of the input. This makes neural networks faster and more stable. Instance normalization prevents contrast shifting by ensuring that the contrast of an input image

is not skewed by images that have a significantly different contrast, which could occur when batch normalization is used. [87, 89] The overall generator network is shown in Figure 3.6. The final segmentation prediction was obtained by applying a sigmoid function to the output of the generator to obtain prediction values in the range of [0,1].

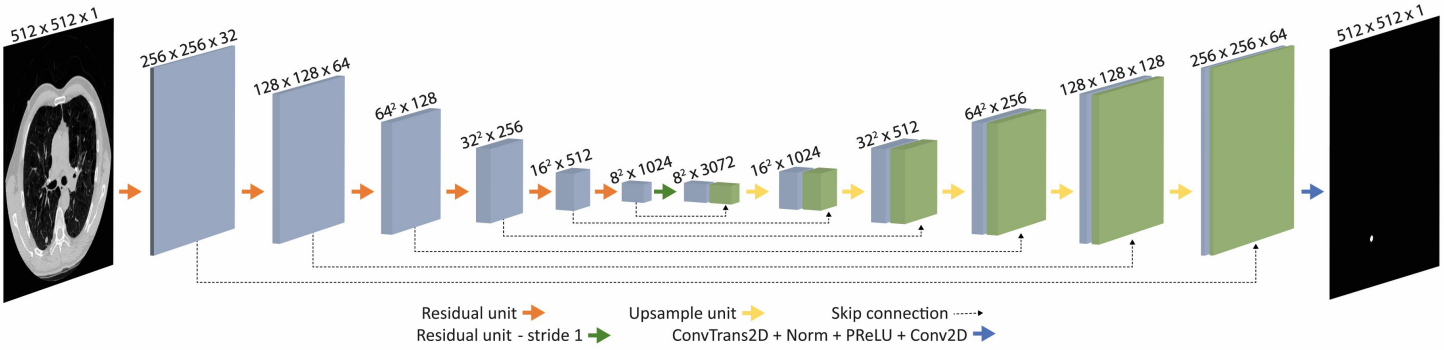


Figure 3.6: Overall structure of the generator network. The size of the feature maps are indicated above each box.

3.2.2 Critic network

Instead of classifying the generated output as real or fake, the proposed discriminator employed a multiscale L1 feature-matching loss between the predicted output and the GT. Therefore, it is referred to as critic network (C) instead of discriminator. The architecture of C was adapted from the SegAN network of Xue et al. [44] and is shown in Figure 3.7. It consists of six convolution layers, which each are set with a stride of two to gradually increase the receptive field and extract feature maps of the nodule at multiple scales. Normalization and Leaky Relu activation were included after each convolution. At each layer of the critic, hierarchical features (pixel-level, low-level and high-level features) were extracted to compute the multiscale L1 loss. Therefore, this loss is able to capture spatial relations between pixels on short-range and long-range. [90] Given a minibatch with N images, the multiscale L1 loss function is defined as:

$$\min_G \max_C \mathcal{L}_{ms}(C,G) = \frac{1}{N}\sum_{n=1}^{N} \ell_{mae}(f_C(y_n \cdot G(y_n)), f_C(y_n \cdot x_n))) \tag{3.2}$$

where y_n is the input image, x_n is the GT, $G(y_n)$ is the predicted segmentation and f_C represents the features extracted by the critic network. ℓ_{mae} is the Mean Absolute Error (MAE), also known as L_1 distance, and is described as:

$$\ell_{mae}(f_C(x), f_C(\hat{x}) = \frac{1}{L}\sum_{i=1}^{L} ||f_C^i(x) - f_C^i(\hat{x})|| \tag{3.3}$$

where L represents the total number of layers (scales) in the critic, and $f_C^i(x)$ the extracted feature map of image x at layer i. At each layer of C, the L1 distance between the features of the GT and the predicted segmentation was calculated. The multiscale L1 loss was formed by averaging the L1 distances of all the layers of the critic. The critic network was given two inputs: the input CT-image masked by the GT, yielding an image that contained only the tumor part, and the input image masked by the generated segmentation. This masking procedure ensured preservation of texture information and facilitates getting effective, rich multiscale features. [71] The generator and the critic were trained alternately. The critic aimed to maximize the multiscale loss, while the generator tried to minimize this same loss function, based on the gradients passed along by the critic. Therefore, both the generator and critic were forced to learn the discriminative features for lung nodules.

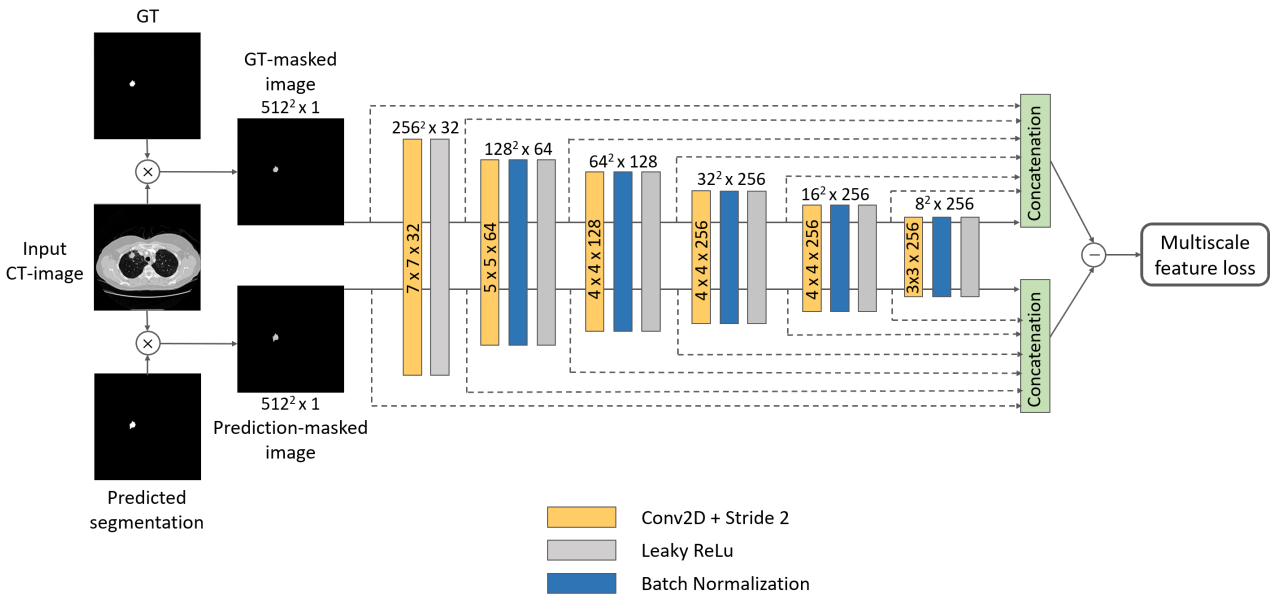


Figure 3.7: Architecture of the discriminator network. Size of each convolution kernel is pointed out in the orange boxes. Resulting size of the feature maps is indicated above the boxes.

3.2.3 Loss functions

In addition to the adversarial loss, a Dice loss was applied to alleviate the class imbalance problem. This loss is defined as [66]:

$$\mathcal{L}_{Dice}(G) = 1 - \frac{2|x \cdot \hat{x}|}{|x| + |\hat{x}|} = 1 - \frac{2\sum_i^I x_i \hat{x}_i}{\sum_i^I x_i + \sum_i^I \hat{x}_i} \tag{3.4}$$

where x and $\hat{x}$ represent the GT and predicted segmentation, respectively, and the sums run over the I pixels of x and $\hat{x}$. The Dice loss can simply be expressed as:

$$\mathcal{L}_{Dice} = 1 - \frac{2TP}{2TP + FP + FN} \tag{3.5}$$

with pixels classified as True Positive (TP), False Positive (FP) and False Negative (FN). Since the background class (TN) is excluded, it cannot overwhelm the smaller segmentation class, which emphasizes the convenience of a Dice loss as a solution to the medical class imbalance problem. Combining the Dice loss with the multiscale L1 loss, the final objective becomes:

$$V(C, G) = \arg\min_G \max_C \mathcal{L}_{ms}(C, G) + \lambda \mathcal{L}_{Dice}(G) \tag{3.6}$$

where λ is a scaling parameter to ensure both loss functions are of the same order of magnitude. As can be seen, the Dice loss only contributes to the generator loss.

Several extensions to the Dice loss exist that could be beneficial for segmentation tasks. The Tversky Index (TI) is a measure that is a generalisation of the Dice coefficient. The Tversky loss, 1-TI, is defined as [91]:

$$\mathcal{L}_{Tversky} = 1 - \frac{TP}{TP + \alpha FP + \beta FN} \tag{3.7}$$

Two parameters are added to the loss, α and β, where $\alpha + \beta = 1$. If $\alpha = \beta = 0.5$, it simplifies to the Dice loss. The significance of these parameters is that one can penalise false negatives or false positives more depending on the value for α and β. In medical applications, false negatives are often less tolerated than false positives. [91] By choosing $\beta > \alpha$, false negatives are penalised

more, which is convenient in highly imbalanced datasets where this additional level of loss function control could yield improved small scale segmentations compared to the normal Dice loss. [91]
A generalisation of the Tversky loss is the Focal Tversky Loss (FTL). This loss gives additional control over how the loss behaves at different values of the Tversky Index, and is given as [92]:

$$\mathcal{L}_{FTL} = (1 - TI)^{\gamma} \tag{3.8}$$

When α and β are chosen to be 0.5, the FTL becomes a Focal Dice Loss (FDL). The parameter γ controls the non-linearity of the loss. When $\gamma < 1$, the gradient of the loss is higher for samples where TI>0.5, which forces the network to focus on such easier examples. This could be useful to incentive learning in the final phase of training, even though TI is nearing convergence. In case of class imbalance, the FTL could become useful when $\gamma > 1$, as this results in higher gradients for samples where TI<0.5. This forces the network to focus on the harder examples, e.g. small scale segmentations, which usually have a low TI score. The non-linearity of the FTL for different values of γ is shown in Figure 3.8.

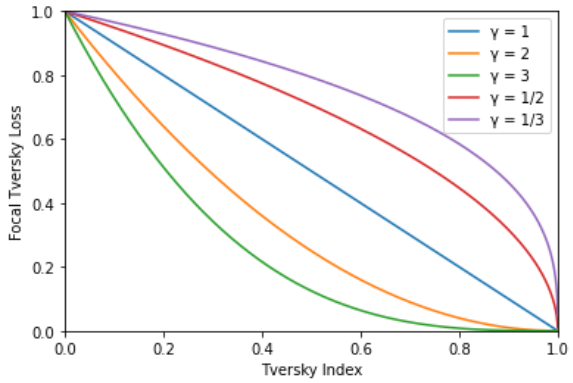


Figure 3.8: FTL as a function of TI with varying values of γ. Image used from [93]

The Dice loss, the Tversky Loss, the FTL and the FDL were all separately tested in the proposed method to determine which loss function was most suitable for our application.

3.2.4 ASPP module

For semantic segmentation, both context information as local information is of great importance. [94] Local context contains detailed information about boundaries and small objects, while global context can locate the larger objects and diminish the impact of similar background and surrounding structures. [70,95] Consequently, as lung nodules have arbitrary sizes, it could be of interest to create a network that is robust to handle these spatial scale variations. Moreover, due to high similarity of intensity values in CT-images, the region to be segmented is easily disturbed by its surroundings, resulting in a loss of semantic information. This could be reduced by applying wider receptive fields to get more valid, surrounding information. [95] Instead of increasing the kernel size, and hence the number of trainable parameters, dilated convolutions can be applied to obtain larger receptive fields. Dilated convolutions, also known as atrous convolutions, add spacing between the elements of the convolution kernels. They thus increase the distance between neighbouring pixels that are considered when computing the value for the center pixel. [96] Therefore, the receptive field can be exponentially increased without a substantial increase of trainable parameters or a loss in resolution, capturing richer information for more accurate segmentation. [95]. In Figure 3.9, the working principle of dilated convolutions with a kernel size of 3×3 and dilation rates of 1,2 and 3 is illustrated. A dilation rate of one is similar to a normal convolution.
Considering a one-dimensional signal, for each location $\mathbf{i}$ on output $\mathbf{y}$ and filter $\mathbf{w}$ with kernelsize $\mathbf{k}$, dilated convolution is applied over the input $\mathbf{x}$. Mathematically, this dilated convolution can be

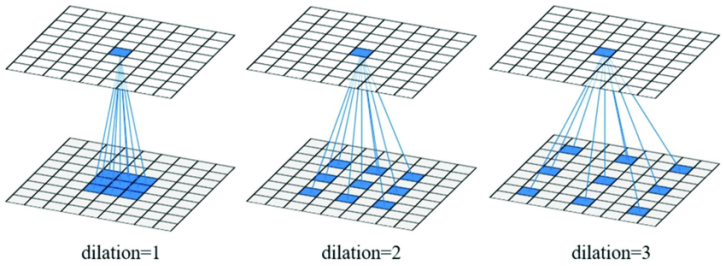


Figure 3.9: Illustrations of dilated convolutions with kernel size 3×3 and various dilation rates. Source: [97]

expressed as follows [98]:

$$\mathbf{y}[\mathbf{i}] = \sum_{\mathbf{k}} \mathbf{x}[\mathbf{i} + r \cdot \mathbf{k}]\mathbf{w}[\mathbf{k}] \tag{3.9}$$

where r is the dilation rate, which corresponds to the stride with which the input signal is sampled. Dilated convolutions with rate r introduce r-1 zeros between consecutive kernel elements, enlarging the kernel size of a $k \times k$ filter to a $(k + (k-1)(r-1)) \times (k + (k-1)(r-1))$ filter. [98]

By applying several dilated convolutions with different dilation rates in parallel, both global and local context of a convolution layer can be extracted at the same time. This approach is called Atrous Spatial Pyramid Pooling (ASPP). [98] The image features extracted from each dilation rate, and thus different sizes receptive fields, are concatenated together to generate a final prediction result. ASPP modules are a promising technique for segmentation of medical images. [48, 70, 95, 99, 100] Wei et al. [70] implemented ASPP modules in the skip connections of a U-Net to capture multi-level contexts without deterioration of resolution for the segmentation of multiscale skin lesions. For the same task, Lei et al [48] implemented a dense dilated convolution block in the bottleneck of their U-Net to obtain fine-grained information and enlarged receptive fields. Xia et al. [95] used ASPP modules after each pooling operation in the encoder of their U-Net for the segmentation of various anatomical structures in CT-images. They showed that this method captures more convenient information to distinguish the target of interest from the similar background. [95]

In the proposed method, ASPP modules containing four parallel dilated convolutions were implemented in the generator network as a possible solution to the influence of the variety in nodule sizes and their similarity to their surroundings and background on the segmentation result. The architecture of the ASPP module can be found in figure 3.10. The dilation rates were set to $r \in [1, 2, 4, 6]$ with kernel sizes $k = [1, 3, 3, 3]$

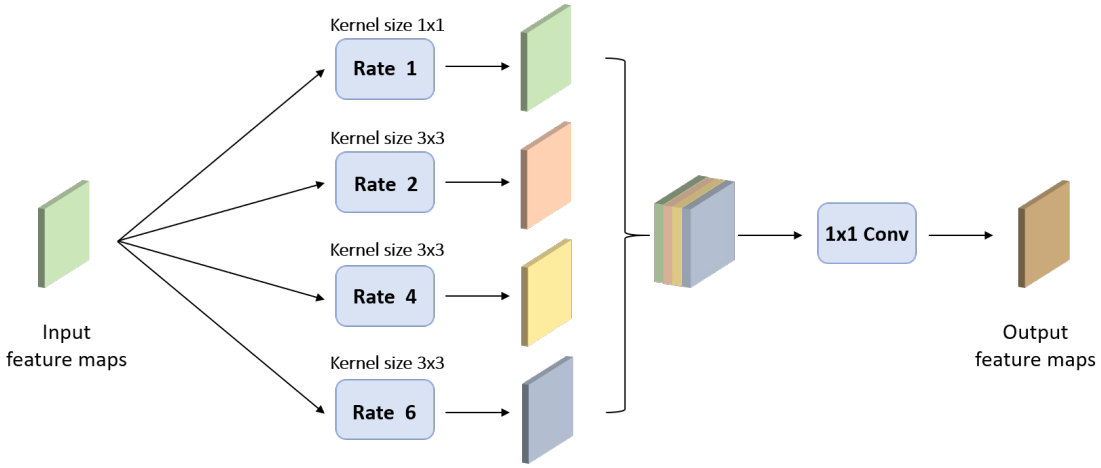


Figure 3.10: The architecture of the ASPP module with four dilation rates.

3.3 Training procedure

The training process of the proposed network was performed by maximizing the critic loss and minimizing the generator loss as a minimax game, equal to the GAN training process described in section 2.1.1. The critic C tries to maximize its loss function:

$$\mathcal{L}_C(\theta_C, \theta_G) = \arg\max_{\theta_C} \mathcal{L}_{ms}(\theta_C, \theta_G) \tag{3.10}$$

whereas generator G tries to minimize its loss function:

$$\mathcal{L}_G(\theta_C, \theta_G) = \arg\min_{\theta_G} \mathcal{L}_{ms}(\theta_C, \theta_G) + \lambda \mathcal{L}_{Dice}(\theta_G) \tag{3.11}$$

This results in the overall objective function described in Equation 3.6. G and C were trained using a stochastic gradient descent optimization. The gradient ascent optimization procedure for the parameters of C is transferred into a gradient descent procedure by multiplying its loss by -1. The entire training process is shown in Algorithm 1.

Algorithm 1 Pseudocode of optimizing the multiscale-L1 GAN

Initialization: N pairs of training data $(y, x)^{(N)}$ extracted from real dataset; hyperparameters: n number of training epochs, k number of mini-batches, mini-batch size m, learning rate α, scaling parameter λ
for n epochs **do**
  for k mini-batches **do**
    Update Critic
    Sample minibatch of m input images $\{y^{(1)}, \ldots, y^{(m)}\}$ from the real dataset;
    Sample minibatch of m corresponding GT masks $\{x^{(1)}, \ldots, x^{(m)}\}$;
    Generate m predicted segmentation outputs $\{\hat{x}^{(1)}, \ldots, \hat{x}^{(m)}\}$, with $\hat{x}_i = G(y_i)$;
    Multiply GT and generated masks with input image: $x_i \cdot y_i$ and $\hat{x}_i \cdot y_i$;
    Compute the loss of the critic network in the current mini-batch:

$$\mathcal{L}_C(\theta_C, \theta_G) = -\frac{1}{m} \sum_{i=1}^{m} \ell_{mae}(f_C(y_i \cdot G(y_i)), f_C(y_i \cdot x_i)))$$

    Update critic network's parameters θ_C by descending its stochastic gradients:

$$\theta_C = \theta_C - \alpha \cdot \nabla_{\theta_C}(\mathcal{L}_C(\theta_C, \theta_G))$$

    Update Generator
    Compute the loss of the generator network in the current mini-batch:

$$\mathcal{L}_G(\theta_C, \theta_G) = \frac{1}{m} \sum_{i=1}^{m} \ell_{mae}(f_C(y_i \cdot G(y_i)), f_C(y_i \cdot x_i))) + \lambda \mathcal{L}_{Dice}(x_i, \hat{x}_i)$$

    Update generator network's parameters θ_G by decreasing its stochastic gradients:

$$\theta_G = \theta_G - \alpha * \nabla_{\theta_G}(\mathcal{L}_G(\theta_C, \theta_G))$$

  end for
end for

Return: Trained and optimized generator network, which can then be applied on the test dataset.

3.4 Experiments

This section provides a description of the various experiments performed in this study, which were designed to be able to answer the sub-questions of this thesis.

3.4.1 Ablation study

An ablation experiment was performed based on the generator network (the U-Net) to verify the effectiveness of each component in the proposed architecture.

1. First, the relevance of the Dice loss for the class imbalance problem was pointed out by comparing its results to a general binary cross-entropy (BCE) loss. Additionally, experiments were realized with the various extensions of the Dice loss to verify which loss function is optimal for our task.

2. Then, the ASPP modules were added to the generator network. The ASPP modules were introduced to three different locations; in the skipconnections, in the bottleneck (i.e. the bottom layer of the U-Net) and after each downsampling layer, to discover which location produced the best segmentation performance and to determine whether this performance improved the results of the generator alone.

3. At last, adversarial training was employed by introducing the critic network to the generator network. First, only adversarial training was enabled, setting λ in Equation 3.6 to zero, to observe the influence of the critic on the performance of the generator. Then, the additional Dice loss was introduced by setting $\lambda > 0$.

The ablation study was performed with the original dataset. Based on this ablation experiment, the optimal combination of modules to obtain the highest possible segmentation performance was observed. If a module turned out to be ineffective, it was not used further in the experiment.

3.4.2 Comparison original and corrected dataset

The best U-Net and GAN-based models, based on the results of the ablation study, were also evaluated on the corrected dataset to determine whether the adaption in GT creation was of significant influence for the segmentation results. The segmentation performance of these models on the corrected dataset were compared to their performance on the original dataset.

3.4.3 Overall performance

From the results of the ablation study and the performance comparison on the corrected dataset, the U-Net based network and GAN based network that showed the best results on the corrected dataset were selected. From these two networks, the overall qualitative and quantitative performance was evaluated and compared. To assess the consistency and robustness of both models, each model was trained and tested three times. Moreover, the models were evaluated on both high and low subtlety-score nodules to assess their robustness for different types of nodules. The test dataset was split into two datasets; one low subtlety set, consisting of 100 images that contained nodules of subtlety score <3.5, and one high subtlety set, consisting of 566 images with nodules that had a subtlety score of ≥ 3.5.
To be able to evaluate the potential increase in training stability of the proposed multiscale-L1 GAN, some more conventional GANs were implemented as well. From literature it was found that WGANs and LSGANs could improve training stability (Section 2.3.1), as well as one-sided label smoothing (Section 2.2.1). Therefore these GANs, along with the basic vanilla GAN, were implemented in this study to compare their stability to that of the GAN with a multiscale-L1 critic.

3.5 Evaluation metrics

For performance evaluation, several evaluation metrics that are relatively common for the evaluation of medical image segmentation were used, namely Dice Similiarity Coefficient (DSC), Hausdorff distance (HD), Precision and Recall. [101] DSC was the main metric to measure the segmentation results.

The DSC is the most used metric in validating medical segmentations. [101] It measures the degree of similarity between the predicted segmentation and the GT, representing the ratio of overlap between both images. Suppose A is the segmentation result of lung nodules predicted by the network, and B is the GT manually segmented by the radiologist. The DSC can be obtained as follows:

$$DSC(A,B) = \frac{2(A \cap B)}{|A| + |B|} = \frac{2TP}{2TP + FP + FN} \tag{3.12}$$

where TP, FP and FN are true positive, false positive and false negative predictions, respectively. In order to calculate the DSC, A needs to be binarized based on a preselected binarization threshold, which is 0.5 by default.

Although DSC can assess the similarity between the predicted segmentation and the GT, it is insensitive to the description of the boundary. As the Hausdorff distance measures the shape similarity and takes pixel localization into consideration, it can provide a valuable complement to the Dice. [7] The HD measures the maximum of the shortest distances from one set of points to another, which indicates the largest segmentation error. [102] Considering two sets of points, A and B, the HD is defined as:

$$HD(A,B) = \max(hd(A,B), hd(B,A)) \tag{3.13}$$

$$hd(A,B) = \max_{a \in A} \min_{b \in B} ||a - b|| \tag{3.14}$$

$$hd(B,A) = \max_{b \in B} \min_{a \in A} ||b - a|| \tag{3.15}$$

For clarification, this is illustrated in Figure 3.11. In image segmentation, the HD is computed between the boundaries of the GT and the generated segmentation. [102]

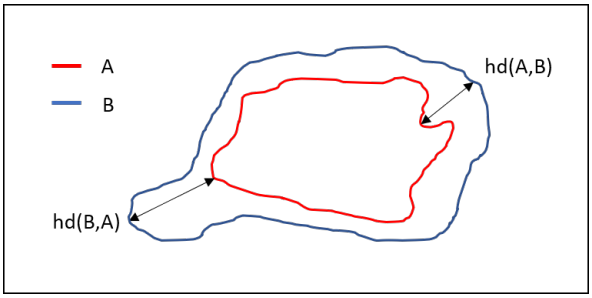


Figure 3.11: A schematic showing the Hausdorff Distance between two point sets A and B

The other performance metrics were used as auxiliary measures. These evaluation criteria are defined by the following formulas:

$$Precision = \frac{TP}{TP + FP} \tag{3.16}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.17}$$

Paired t-tests were performed to analyse whether observed differences in DSC score between two models were significant (at the level of 95%). A paired t-test is a statistical test that compares the means of two related groups and determines whether the mean difference between paired samples of these groups is zero (i.e. the null hypothesis). [103] In this study, paired samples correspond to the same set of input images processed by two different models. The t-test is of the form *samples mean difference / samples standard error* and returns the p-value, p, which represents the probability of finding the observed differences when the null hypothesis is true. [103] A difference was assumed to be significant if $p < 0.05$. In the ablation study, these tests were used to determine if a module significantly improved the performance of the U-Net.

3.6 Implementation details

All networks were implemented in Python 3.8 with the usage of the PyTorch framework and trained on one NVIDIA A40 GPU using CUDA 11.0 for accelerated training. The U-Net was implemented using MONAI, an open-source, PyTorch-based framework for deep learning in healthcare imaging. [104] For the critic, the architecture of the SegAN critic was used and adapted to our purpose. All models were initialized from a Gaussian distribution $\mathcal{N}(0, 0.02)$ and optimized using Adam with $\beta 1 = 0.5$ and $\beta 2 = 0.999$. A learning rate schedule was implemented that decayed the learning rate by 0.5 every 25 epochs. The hyperparameters for both U-Net and GAN training were tuned separately to find the optimal versions of both. For the U-Net alone, the batch size was set to 20 and the initial learning rate was 0.0001. The network was trained for up to 150 epochs.
In the proposed GAN-based method, the learning rates for both the generator network and the critic network were initialized as 0.0002 and decayed following the learning rate schedule with a minimum learning rate of 10^{-8}. A batch size of 12 was used, which was the maximum possible size due to GPU memory limitations, and the network was trained for up to 250 epochs. λ was set to 0.2 to match the Dice loss' order of magnitude to the adversarial loss' magnitude. The weights of the critic were clamped in the range of [-0.05 0.05] every time they were updated through gradient descent to reduce the likelihood of gradient exploding. In both the U-Net as and the critic of the GAN-model, dropout with ratio 0.2 was enabled to improve the generalization capability of the networks and to prevent overfitting.
After each training epoch, the model was tested on the validation set and evaluated by the Dice score. After the maximum amount of training epochs, the Dice score on the validation set had become stable. The model with the best validation Dice score was saved and its performance was evaluated on the independent test set.

4 Results

This chapter shows the results of the experiments as described in Section 3.4. The outcome of the ablation study is demonstrated in Section 4.1 and a comparison between the original and corrected dataset is given in Section 4.2. The results of the best U-Net are compared to the results of the best GAN-based method in Section 4.3 to evaluate the value of adversarial training in addition to a U-Net. At last, the performance of the proposed models is compared to state-of-the-art segmentation approaches in Section 4.3.4.

4.1 Ablation study

The overall results of the ablation study are shown in Table 3. In the following paragraphs, the results are described more extensively for each module.

4.1.1 Loss functions

The proposed U-Net was trained with a BCE loss, a Dice loss and multiple extensions of the Dice loss: two U-Nets with focal Dice loss ($\gamma = 0.5$ and $\gamma = 2.0$), two U-Nets with Tversky loss ($\alpha = 0.7$, $\beta = 0.3$ and $\alpha = 0.3$, $\beta = 0.7$), and two U-nets with focal Tversky loss ($\alpha = 0.3$, $\beta = 0.7$, $\gamma = 0.5$ and $\alpha = 0.3$, $\beta = 0.7$, $\gamma = 2.0$). For the U-Net with a BCE loss, a best validation Dice score of 0.0012 was obtained at the first epoch, after which it almost immediately dropped to zero.

The results of the experiment with the other loss functions are visualized in Figure 4.1. The Dice scores of all models were rather compatible. The main differences occurred in the precision and recall scores. In Figure 4.1, the U-Net with FDL ($\gamma = 0.5$) achieved the best performance on the Dice score. However, this improvement compared to the U-Net with Dice loss was not significant (p=0.477). The quantitative performance results of these two models are shown in Table 7. For more details about the other loss functions' performance and their precision-recall curves, the reader is referred to Appendix B.1

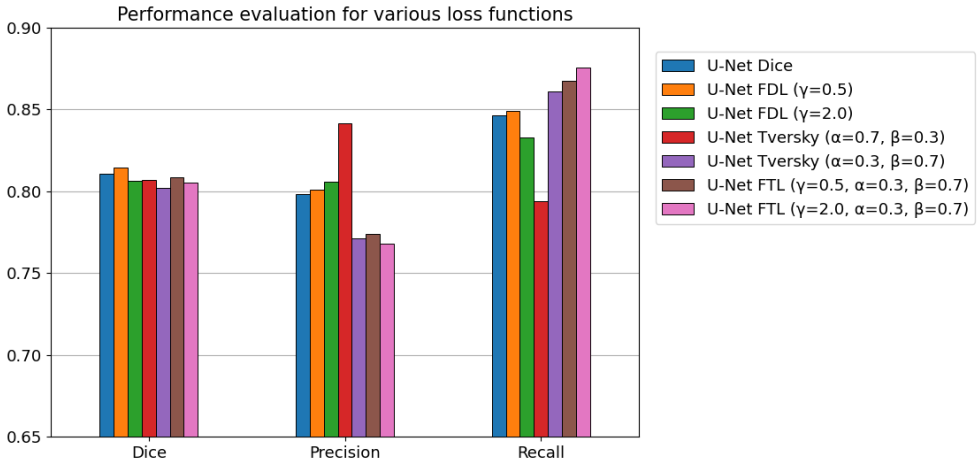


Figure 4.1: Performance of various loss functions

4.1.2 ASPP module

ASPP modules were implemented at three different locations in the U-Net and the output was compared to the U-Net without ASPP modules. Table 2 shows the experimental results. Figure 4.2 illustrates the results in a more intuitive way by bar plots. The models with ASPP modules implemented in the skipconnections and the downsampling path yielded significantly worse performance

Table 1: Quantitative performance results of the U-Net with Dice loss and the U-Net with focal Dice loss. The values are given as mean ± standard deviation.

Loss	DSC	Precision	Recall
Dice	0.811 ± 0.220	0.798 ± 0.238	0.846 ± 0.216
FDL ($\gamma = 0.5$)	0.814 ± 0.226	0.801 ± 0.239	0.849 ± 0.228

than the U-Net without ASPP modules (both $p < 0.001$). The U-net with an ASPP module in the bottleneck achieved comparable results to the U-Net without ASPP. Based on the paired t-test, no significant improvement was obtained by the addition of an ASPP module to the bottleneck of the U-Net ($p = 0.836$).
Considering the standpoint of scale-space theory, the ASPP modules were also tested with Gaussian convolutions succeeding the parallel dilated convolutions. More details on this experiment and its results can be found in Appendix B.2

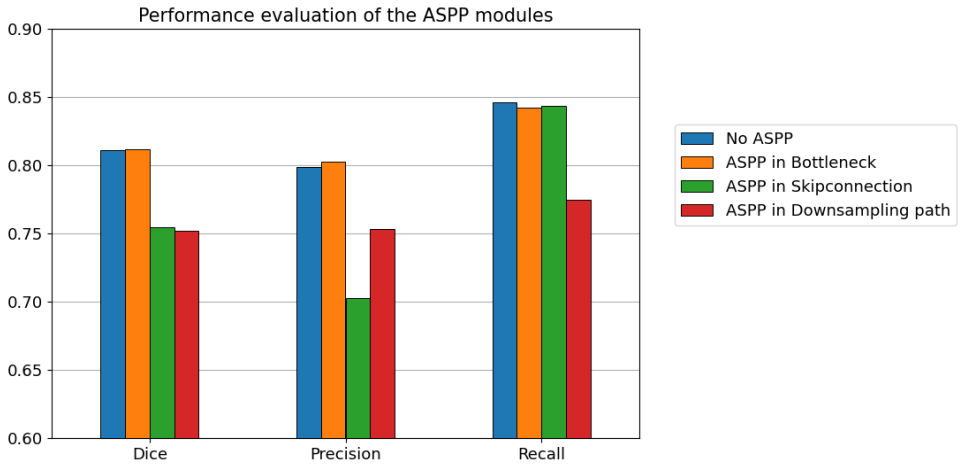


Figure 4.2: Performance of U-Net with ASPP modules implemented at various locations.

Table 2: Performance results of the U-Net model with ASPP modules implemented at different locations. Values are given as mean ± standard deviation.

Location ASPP	DSC	Precision	Recall
No ASPP	0.811 ± 0.220	0.798 ± 0.238	0.846 ± 0.216
Bottleneck	0.812 ± 0.237	0.803 ± 0.251	0.842 ± 0.236
Skipconnection	0.755 ± 0.239	0.702 ± 0.247	0.843 ± 0.249
Downsample path	0.752 ± 0.271	0.754 ± 0.290	0.775 ± 0.278

4.1.3 Critic network

The effect of adding a multiscale-L1 critic to the U-Net, i.e. conversion into a GAN, is shown in Table 3. Adversarial training showed worse results compared to applying the generator alone. The GAN itself produced many non-overlapping results between the predicted segmentation and the GT, in which the network often predicted no tumor pixels at all. A histogram of the Dice scores

predicted by this GAN is shown in Appendix C. With an additional Dice loss, the performance of the GAN increased considerably. Nevertheless, it still obtained significantly inferior performance compared to the U-Net without a critic ($p < 0.001$).

4.1.4 Overall results ablation study

The overall results of the ablation study are summarized in Table 3. For a visual representation of the DSC score distributions of the methods, in the form of boxplots, the reader is referred to Appendix B.

Table 3: Overall results of the ablation study. Outputs are given as mean ± standard deviation. P-values are based on a paired t-test between the DSC values of the method and the U-Net.

Method	DSC	Precision	Recall	p
U-Net	0.811 ± 0.220	0.798 ± 0.238	0.846 ± 0.216	-
U-Net + FDL	0.814 ± 0.226	0.801 ± 0.239	0.849 ± 0.228	0.477
U-Net + ASPP	0.812 ± 0.237	0.803 ± 0.251	0.842 ± 0.236	0.836
GAN[1]	0.441 ± 0.402	0.460± 0.421	0.465 ± 0.431	<0.001
GAN[1] + Dice	0.772 ± 0.268	0.780 ± 0.280	0.788 ± 0.274	<0.001

[1] GAN = U-Net + Critic

4.2 Comparison original and corrected dataset

In Table 4, a comparison was made between the performance of the U-Net and GANs on the original and corrected dataset. The values represent the mean Dice score ± the standard deviation. All models showed improved performance on the corrected dataset.

Table 4: Comparision between performance of models trained on the original dataset and the corrected dataset. The values represent the mean Dice score and the standard deviation. Values in bold represent the best U-Net outcome and the best GAN outcome.

Method	Original data	Corrected data
U-Net	0.811 ± 0.220	**0.845 ± 0.163**
GAN	0.441 ± 0.402	0.472 ± 0.391
GAN+Dice	0.772 ± 0.268	**0.808 ± 0.222**

4.3 Overall performance

Based on the results of the ablation study and Table 4, the best U-Net and the best GAN-based model were retrieved, i.e. the U-Net with a Dice loss and the GAN with an additional Dice loss, respectively. The U-Net with Dice loss was chosen as the other modules (ASPP, FDL) did not show significant improvement. The overall performance of the two models can be found in Section 4.3.1 and the performance on high and low subtlety-score nodules separately are shown in Section 4.3.2. The segmentation results are visualized in Section 4.3.3. Lastly, a comparison with the results of other existing methods is made in Section 4.3.4 to depict the efficiency of the proposed models compared to the state-of-the-art. For more details on the results of the GAN network without the additional Dice loss, the reader is referred to Appendix C.

4.3.1 Performance evaluation

The evaluation metric scores of both models are depicted in Table 5. The models were trained three times and the values represent the mean ± standard deviation over three runs. The HD is generally sensitive to outliers, which was substantiated by the boxplots in Figure 4.3. Most HD values had a value around two pixels. However, many outliers existed due to non-overlapping segmentation predictions, which had a maximum value of 343.3 for the U-Net and 492.9 pixels for the GAN. These extremely large outlier values can give a distorted view of the HD performance. Therefore, the mean HD was calculated without the outlier values as well. In Table 5, the HD is represented by two values: HD+ simply represents the mean HD value of all output predictions, while HD− corresponds to the mean HD value after removal of outliers.
The average time consumption for training the U-Net was 4.5 hours and for the GAN 16 hours. Both models showed a stable training course during training. The graphs of the models' training progress can be found in Appendix D.

Table 5: Performance evaluation of the best U-Net and GAN-based model. Values are given as mean ± std over three runs. HD+ represents the mean HD value with outliers included, HD- is the mean HD value without outliers. HD is given in pixels.

Method	DSC	Precision	Recall	HD+ / HD-
Best U-Net	0.844 ± 0.002	0.851 ± 0.005	0.857 ± 0.009	11.5 / 2.0
Best GAN	0.807 ± 0.007	0.820 ± 0.008	0.818 ± 0.006	26.9 / 2.3

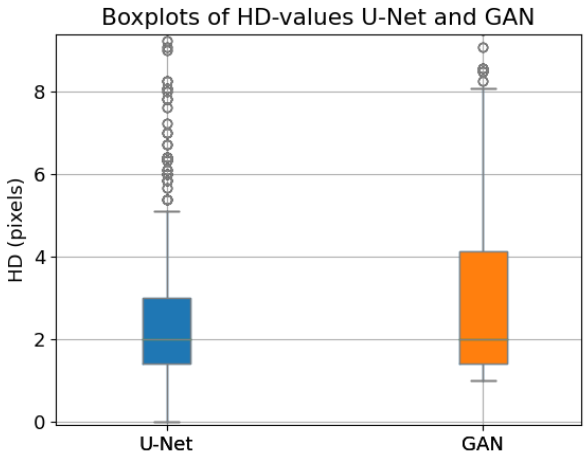
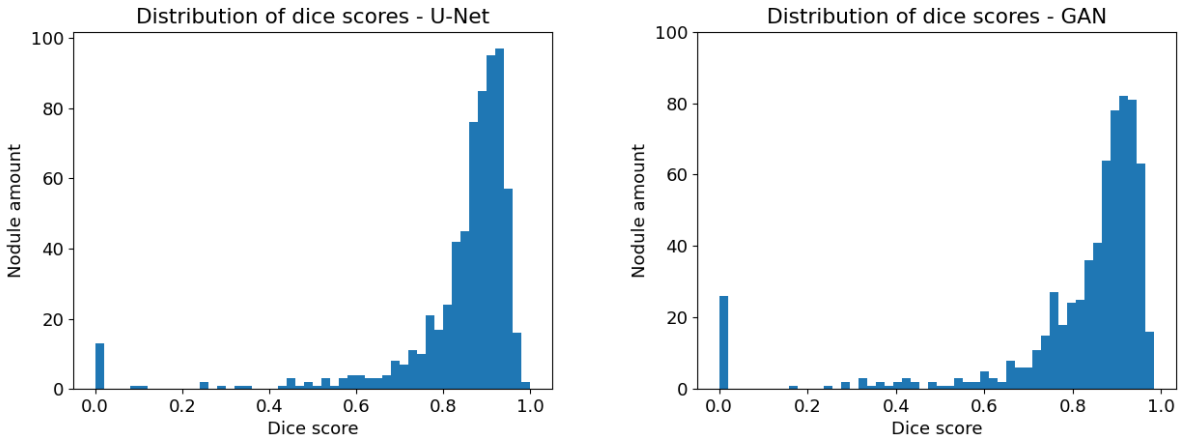


Figure 4.3: Boxplot of HD values for both the U-Net and GAN model. Graph is zoomed in on relevant part of the boxplot, removing most of the outliers.

A histogram of all Dice values is shown in Figure 4.4 for a more complete evaluation of the output of the U-net and GAN on the test set. It can be seen that for both models, most nodules obtained a Dice score greater than 0.8. The distributions are somewhat left-skewed with a long tail to the left. For some nodules, the prediction did not overlap at all with the GT, i.e. failure cases. This applied to thirteen predictions of the U-net and twenty-six of the GAN. Removing these failure cases from the data led to an elevated DSC of 0.862 ± 0.113 for the U-Net and 0.853 ± 0.118 for the GAN.



(a) DSC values of U-Net on test set.

(b) DSC values of GAN on test set.

Figure 4.4: Distribution of Dice values on test for both models.

Previous results were all based on a default binarization threshold of 0.5 to classify the pixels of the networks' output into tumor or background pixels to obtain the final segmentation result. However, the choice of threshold value could alter the outcome of the three evaluation metrics. This is shown in Figure 4.5. For one U-Net run, the Dice, precision and recall scores were determined at 25 threshold values between zero and one. It appeared that the Dice score remained more or less constant for most of the thresholds, while the precision and recall scores varied considerably.

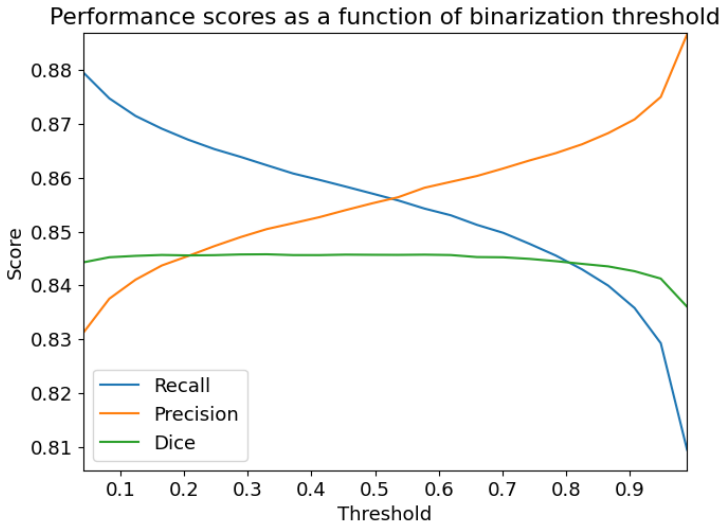


Figure 4.5: Recall, precision and Dice scores as af function of binarization thresholds.

4.3.2 Performance on high and low subtlety-score nodules

Figure 4.6 illustrates the performance differences that occurred between testing on the high and low subtlety-score nodule subsets. The U-Net obtained a mean Dice score of 0.857 on high subtlety-score nodules and 0.768 on low subtlety-score nodules, whereas the GAN achieved a mean Dice score of 0.826 and 0.696 respectively. Both models performed better on nodules of high subtlety-score than on low subtlety-score nodules, although this difference was considerably smaller for the U-Net than for the GAN.

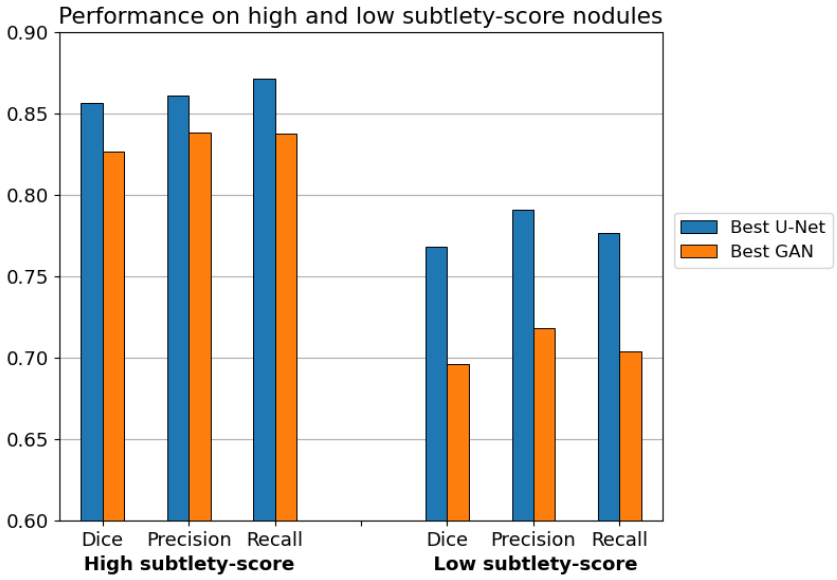


Figure 4.6: Performance of the best U-Net and GAN on the high and low subtlety-score sub-datasets.

4.3.3 Visual results

Figure 4.7 shows two examples of high scoring segmentations on high subtlety-score nodules. The GT (white pixels) were overlayed by the predicted segmentation (red pixels) to accurately see which pixels were classified (in)correctly, resulting in pink pixels at locations where both the GT and the predicted segmentation contained a nodule pixel. On both images, the models reached a Dice score of > 0.95.

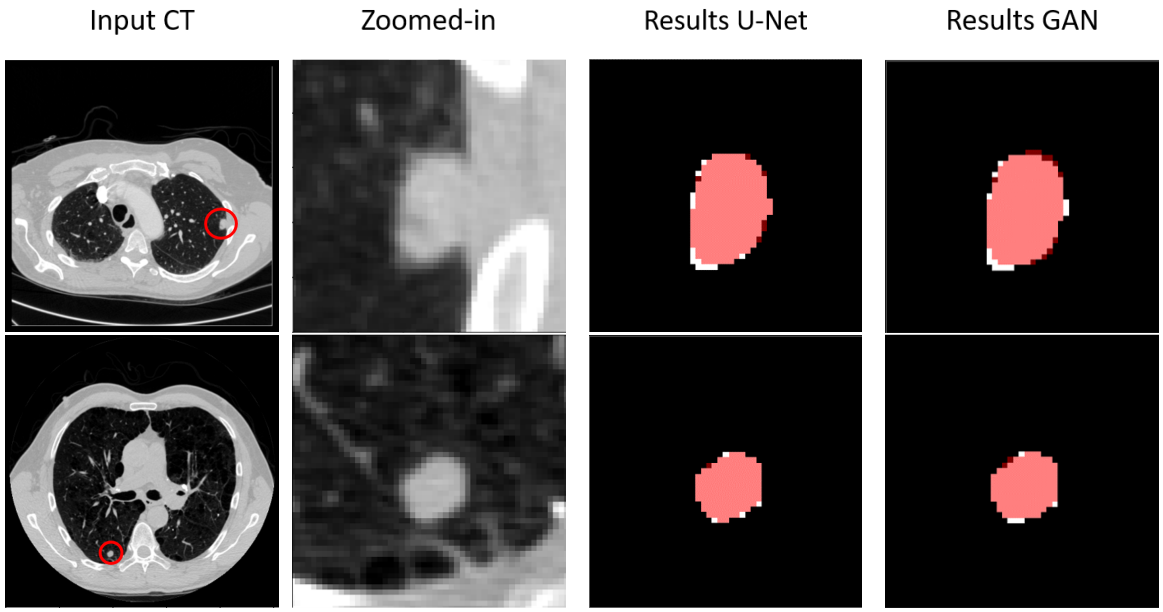


Figure 4.7: Good segmentation results of high subtlety-score nodules produced by the U-net and GAN. Typical examples of input image with high subtlety-score nodule indicated by a red circle (left column), input image zoomed in on nodule (2nd column), segmentation performance U-Net (3rd column) and segmentation performance of the GAN (right column). The GT and the prediction of the network are superimposed to visualize the segmentation accuracy: GT in white and prediction in red. All images, except the input image, have a size of 50×50 pixels.

The qualitative results of segmentation on low subtlety-score nodules are illustrated with two examples in Figure 4.8. As can be seen, these nodules are very similar to the background or surrounding structures and are therefore more difficult to distinguish than the high subtlety-score nodules. Nevertheless, both models showed good performance on these nodules as well. On the first input image, the U-Net and GAN achieved a DSC value of 0.936 and 0.931, respectively. A slightly lower DSC of 0.829 and 0.812, by the U-Net and GAN respectively, was obtained on the second input image.

Although the models performed similarly on the examples shown in Figure 4.7 and 4.8, there were also cases in which the U-Net performed well, while the GAN did not produce reliable predictions or none at all. An example is shown in Figure 4.9. The U-Net obtained a Dice score of 0.791, while the GAN could not distinguish any nodule pixels at all and thus obtained a Dice score of 0.

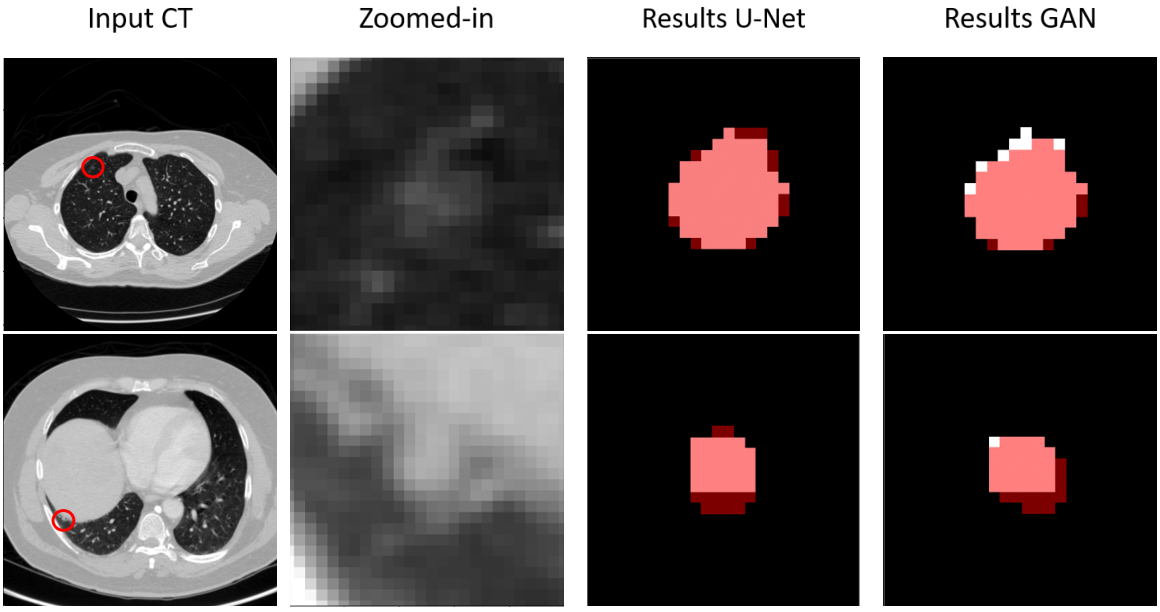


Figure 4.8: Visualisation of the segmentation performance of the U-net and the GAN on low subtlety-score nodules. Typical examples of input CT-image with low subtlety-score nodule indicated by a red circle (left column), input image zoomed in on nodule (2nd column), segmentation performance U-Net (3rd column) and segmentation performance of the GAN (right column). White pixels represent the GT, while red pixels correspond to the predicted segmentation. All images, except the input image, have a size of 25×25 pixels.

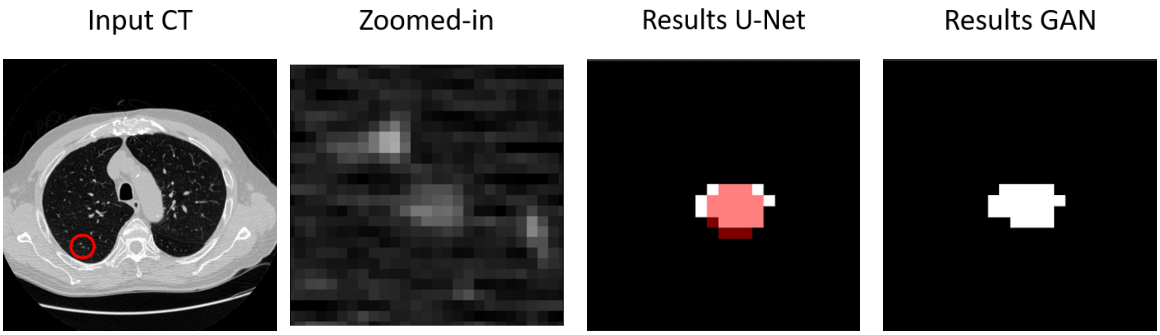


Figure 4.9: Example of well-performing U-Net on a low subtlety-score nodule (DSC=0.7907), while the GAN showed no performance at all. GT pixels are marked as white, while predicted pixels are shown as red.

At last, there were also some cases in which both models did not perform well. These were cases (i) in which a low DSC value was yielded, (ii) in which the models did not predict any tumor pixels at all, resulting in a DSC of 0 and a HD of infinitive, or (iii) cases in which there was partial overlap, i.e. when the model missed one out of two nodules or hallucinated a nodule. Examples of these

cases are demonstrated in Figure 4.10.

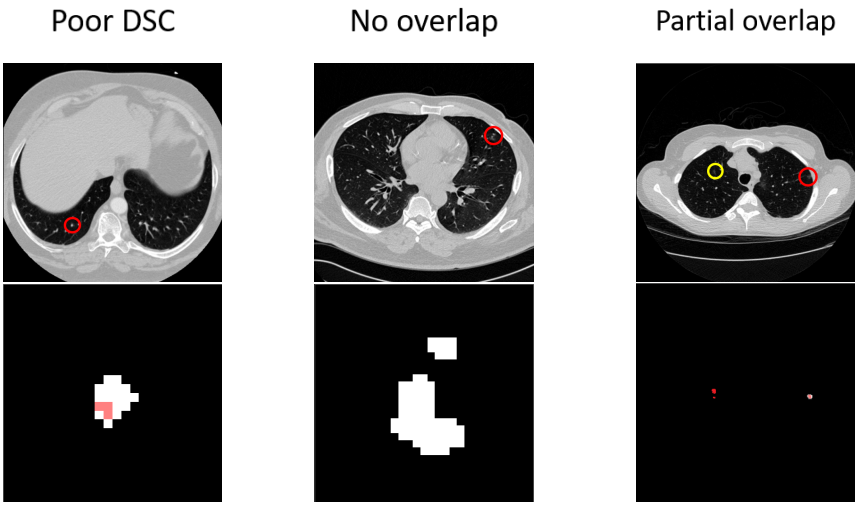


Figure 4.10: Visual examples of segmentation results on which the methods failed to accurately segment the nodule. Left column shows an example of a low Dice score segmentation result (DSC = 0.273). Middle column shows an example in which the model predicted no nodule-pixels at all. Right column shows an example in which the model predicted two nodules (indicated with yellow and red circle), while the GT only contained one nodule (red circle).

4.3.4 Comparison state-of-the-art methods

To evaluate the performance of the best U-Net and GAN in this work, the final results were compared with several other models that were developed in recent years for the same task. The results can be found in Table 6. The U-Net implemented in this thesis showed improved performance on the Dice score and the precision, but inferior results on the recall, compared to other existing methods. In contrast to this study, most previous works have applied patch-based segmentation methods instead of segmenting nodules from entire CT-images. A comparison with these patch-based state-of-the-art networks was made in Appendix E.
Table 6 depicts that the proposed GAN could not surpass the patch-based AUGAN on performance, but showed a significant improvement compared to the CTumorGAN.

Table 6: The quantitative segmentation results of the best U-Net and GAN-based model compared to literature. The values are given as the mean score ± the standard deviation of the scores in percentage. Values in bold represent the best U-Net outcome and the best GAN outcome.

Method	DSC (%)	Precision (%)	Recall (%)
U-Net			
RUN (Lan et al.,2018) [37]	71.9	-	90.9
U-DET (Keetha et al.,2020) [38]	82.8 ± 11.7	78.9 ± 17.5	**92.2 ± 14.1**
Proposed U-Net	**84.4 ± 17.0**	**85.1 ± 18.6**	85.7 ± 17.6
GAN			
CTumorGAN (Pang et al.,2020) [61]	71.1	77.3	70.4
AUGAN[1] (Shi et al.,2020) [7]	**85.3**	**85.6**	**86.0**
Proposed GAN	80.7 ± 21.8	82.0 ± 23.2	81.8 ± 22.5

[1] Patch-based method

4.3.5 Additional results

Besides the proposed multiscale-L1 GAN, experiments were performed with a vanilla GAN, a LS-GAN and a WGAN, whether or not in combination with one-sided label smoothing, for the task of lung nodule segmentation to investigate their stability. Furthermore, the openly-available Pix2Pix algorithm was also applied to our dataset. All of these GANs did not show a stable training course and could not produce accurate segmentation results. Examples of unstable training results are illustrated in Appendix G.
In addition to the multiscale L1 loss, experiments with a multiscale-L2 critic were performed. Adversarial training using a multiscale L2 loss did not converge and seemed to be unstable. A more extensive description of the multiscale L2 loss and its training results are provided in Appendix F.

5 Discussion

Segmentation of lung nodules in CT-images is a critical step in many lung cancer procedures. [6] To assist radiologists in this process, computer-assisted segmentation systems could be a promising tool. Although there have been many related studies before, accurate nodule segmentation on entire 2D CT-images remains a challenging goal. This study aimed to explore the feasibility of applying GANs to automatically segment lung nodules from entire images. The proposed network architecture was designed to address three challenges: (i) the class imbalance, (ii) the variability in lung nodule appearance, and (iii) the GAN training instability. To alleviate the above challenges, the following measures were adopted: a Dice loss to counteract the class imbalance, ASPP modules to reduce the influence of appearance variability, and a multiscale-L1 critic for training stability. The added value of each module to the generator network was evaluated in the ablation study to determine whether this module should be used. Interpretation of the results of the individual modules and the overall performance are discussed in Section 5.1. The limitations of the study are described in Section 5.2.

5.1 Interpretation of the results

First, the proposed modules and their potential to alleviate the segmentation challenges will be evaluated based on the results of the ablation study. Then, the overall performance of the U-Net and the GAN will be discussed.

5.1.1 Evaluation of the modules and their impact on the segmentation challenges

Loss functions Due to severe class imbalance, the results of the U-Net with a BCE loss instead of a Dice loss were not satisfactory. The network predicted all pixels as background, as this results in an extremely low loss. Therefore, the BCE loss was not suitable for our task. For both the U-Net and the GAN, the Dice loss appeared to be an essential addition to obtain satisfactory performance results. The Dice loss prevented the networks from predicting all pixels as the most common class, i.e. the background. Hence, it addresses the class imbalance problem, which was in line with our expectation. The various extensions of the Dice loss did not show significant difference in Dice score. In this study, the main evaluation metric was the Dice score. If the focus were to be placed on the recall or precision, different conclusions could have been drawn. Substantial differences between the loss functions did occur in these metrics due to distinct focus on false positive and false negative predictions.

ASPP module As was shown in Table 2, the ASPP modules obtained worse or comparable results to the U-Net without ASPP module. Out of the three locations, the ASPP in the bottleneck achieved the best performance. However, it is questionable to what extent the ASPP module truly has influence when implemented on this location, as the feature maps size at the bottleneck is only 8×8. This implies that the 3×3 convolution with dilation rate 2 covers almost the whole feature maps in the bottleneck. In addition, the 3×3 convolutions with dilation rates 4 and 6 overstep the boundary of the feature maps, causing it to degrade to a 2×2 or even a 1×1 convolution. This makes it difficult to effectively access long-range features, which was also described by Shen et al. [105] Experiments with different dilation rates were performed, but dilation rates $r \in [1, 2, 4, 6]$ showed best performance.
It was assumed that the ASPP modules could contribute to addressing the variability in nodule size and appearance by capturing richer features and extracting both local and global context at each scale. Among others, Xia et al. [95], Lei et al. [48] and Wei et al. [70] have shown promising segmentation performance on medical images by implementing ASPP modules to reduce the effect of scale variations and resemblance to background structures. However, these studies were performed on patches of lesions or entire organs, whereas our study was performed on small lesions in entire CT-images. A possible explanation for the poor performing ASPP modules in our study could be

that, due to the extremely small size of the nodules compared to the background, no benefit is gained from more global context for this specific task.

Critic A multiscale L1 adversarial loss was applied to increase the stability of the GAN. The proposed GAN showed a stable course during training (see Appendix D). For comparison, we also tried to implement a GAN with the conventional single scalar real/fake loss. However, we did not manage to obtain acceptable segmentation predictions with this GAN due to instability problems. Multiple possible solutions, explained in Section 2.2.1 and 2.3.2, were tried to stabilize training: WGAN, LSGAN, one-sided label smoothing and parameter clamping. Unfortunately, these methods did not solve the instability problem. Moreover, the widely-used Pix2Pix framework was applied on our dataset. This image-to-image translation method could not give stable and accurate predictions as well. Pang et al. [61] also experienced that classical image-to-image translation networks, such as cGAN and Pix2Pix, produced unstable and unacceptable predictions for the segmentation of lung nodules on entire CT-images as input. Training on entire images instead of patches contributes to the GAN training instability, as this increases the number of parameters and could create an overfitting problem. [106] All findings substantiate our expectation that the proposed critic would lead to a more stable and effective GAN-model than the conventional discriminator for the segmentation of lung nodules. Xue et al. [44] showed similar results for brain tumor segmentation. They conclude that the main reason of unstable training is that the conventional adversarial loss is based on a single scalar output. For segmentation tasks, this real/fake classification task could be too easy for the discriminator, leading to insufficient gradient feedback to improve the generator. [44].
In this study, the L1 norm was adopted in the multiscale feature loss function. Experiments were also performed with the L2 norm. As is shown in Appendix F, adversarial training using a multiscale L2 loss did not converge and tended to be unstable. Therefore, the L1 loss was chosen to be used in this study instead of a L2 loss. Xue et al. [44] experienced the same issue and attributed it to the speculation that L1 is less sensitive to outliers than L2.

5.1.2 Overall performance

GAN One of the goals of this study was to examine the usability of GANs compared to the U-Net for lung nodule segmentation. It was assumed that the addition of a critic would urge the generator to produce more accurate segmentations. This would lead to improved performance of the GAN over the generator alone (the U-Net). Although the proposed GAN with an additional Dice loss showed promising results, it did not outperform the U-Net alone on any evaluation metric (see Table 5). Based on these results, it cannot be stated that applying a critic to the generator network is of added value for our task, it only seemed to confuse the generator. Our speculation is that the segmentation of relatively small lesions compared to the background may need more pixel-level focus. Accordingly, small sized lung nodules on entire CT-images could benefit more from a pixel-level loss than the more general loss of a GAN. This is also substantiated by the necessity of using an additional Dice loss to adversarial training. Purely adversarial training obtained significantly lower performance (DSC=0.472) than adversarial training in combination with a Dice loss (DSC=0.808). The indispensability of the Dice loss emphasizes the need of a pixel-level focus and reduced class-imbalance. To validate these claims, it would be interesting to see whether the proposed GAN would outperform the U-Net on patches of nodules instead of entire images, in which the nodules are more substantial and the background is of lesser influence. Shi et al. [7] showed that their GAN did outperfom the U-Net on lung nodule patches. Xue et al. [44] also experienced that their SegAN showed inferior performance for relatively small brain tumor regions.
Out of the three GAN-based models, the AUGAN developed by Shi et al. [7] showed best performance on all evaluation metrics (See Table 6). However, this model was trained and tested on 128×128-patches of lung nodules instead of entire 512×512-images. To the best of our knowledge, the CTumorGAN developed by Pang et al. [61] is the only other GAN-based method for lung nodule segmentation that uses entire 2D CT-images. The proposed GAN with additional Dice loss achieved

improved performance compared to the CTumorGAN. However, the CTumorGAN was applied on the NSCLC-dataset instead of the LIDC-IDRI dataset. For a complete comparison, our GAN should be trained and tested on the NCSLC-dataset to draw fair conclusions about the performance of our GAN.

U-Net The proposed U-Net was able to obtain accurate segmentation of lung nodules. In comparison to other existing methods for lung nodule segmentation on entire CT-images, the proposed U-Net has gained state-of-the-art performance on the Dice value and precision score (Table 6). A possible explanation for its improved performance compared to the U-DET is that the U-DET uses max-pooling operations instead of strided convolutions and does not include residual units. Although our U-Net's performance on the recall score seems inferior to the other methods, its value can be increased by changing the binarization threshold, as was shown in Figure 4.5. It appeared that the Dice score was not sensitive to the threshold of binarization, whereas the precision and recall scores dependeded considerably on the choice of binarization threshold. Lower thresholds are favourable for the recall, while higher thresholds are more convenient for the precision. Depending on the application, it can be weighed whether to focus on the precision or recall and choosing the threshold value accordingly.
It is worth noting that the majority of previous studies have proposed patch-based lung nodule segmentation methods. In this study, it is demonstrated that the proposed U-Net can achieve comparable, and possibly even better results than the patch-based methods (See Appendix E).
Although the proposed U-Net showed promising results, there were still some failing cases (Figure 4.10). The exact reason why the network sometimes fails to produce accurate segmentations is unknown, though the failing cases often comprise nodules that are visually hard to distinguish from the background or surrounding structures. To solve this problem, additional images similar to these situations might be required to enhance the prediction on these cases as well.

Robustness The robustness of the models was analysed by comparing their performance on high and low subtlety-score nodules. Figure 4.6 depicts that the models' potential for segmentation depends on the type of nodule. The DSC, precision and recall of the segmentation results on the high subtlety-score nodules exceeded those of the low subtlety-score nodules. The difference in DSC value between high and low subtlety-score nodules was 0.089 for the U-Net and 0.130 for the GAN. This implies that the U-Net is more robust to various types of nodules than the GAN. Other researchers analysed their model's robustness by only comparing its performance on small and large nodules, instead of taking into account the subtlety-score. The AUGAN showed a DSC score difference of about 0.1 between small and large nodules, which is comparable to the values we found. [7] The robustness of the U-DET was analysed by comparing its performance on Attached and Non-Attached nodules and large ($\geq 6mm$) vs. small ($<$6mm) nodules, obtaining a DSC score difference of only 0.0129 and 0.01 respectively. [38] Therefore, the U-DET showed great potential for robust segmentation and is less dependent upon the type of nodule than the proposed U-Net. However, they do not include all low subtlety-score nodules in their robustness analysis, such as GGO nodules, which were included in our subsets.
The consistency of the models was analyzed by performing multiple runs of the same model. Based on the standard deviation values given in Table 5, both models showed great consistency over multiple runs. Especially the U-Net is extremely consistent on the Dice score, achieving a standard deviation of only 0.002. This suggests that the results are reproducible and consistent.

Statistics The performance values were given as a mean $\pm$ standard deviation of all the observations. This also included failure cases that yielded no overlap with the GT at all. These failure cases were considered in the calculation of the mean performance since they do indicate how well the networks perform. A different way of dealing with these failure cases could be to first determine whether a nodule is detected prior to calculating the segmentation performance. This would split the performance into a detection performance and a segmentation performance. As was shown in

Section 4.3.1, this would result in an increased mean Dice score and a reduced standard deviation. Since the distributions of the performance values were left-skewed with a long tail to the left, large standard deviation values occurred. This explains the existence of standard deviation values that exceeded the upper value of 1 when added to the mean value.
In order to reach an appropriate statistical conclusion about the outcome of the paired t-tests, several conditions must be satisfied: the two groups for comparison must be independently sampled from the same population and the difference between each pair is assumed to be normally distributed without outliers. [103] Although the DSC score distributions were left-skewed and not completely normally distributed, the differences between two distributions did follow a normal distribution. However, outliers still existed. Moreover, strictly independence cannot be assured since multiple slices from the same patient may have appeared in the data. The degree of deviation from the assumptions affects the quality of the statistical conclusions. Nonetheless, t-tests are known to be robust, and in practice, real-world data rarely fully meets the constraints of statistical models. [103]

5.2 Limitations

A number of limitations of the study can be identified.
First of all, the ablation study was performed based on the original dataset. Towards the end of the study, it appeared that this dataset was not completely correct. A new, corrected dataset was created and applied on most of the main models. However, due to time constraints, we did not redo all the experiments of the ablation study. It was assumed that the relations found in the ablation study on the original dataset would apply equally on the corrected dataset. All experiments should be re-performed with the corrected dataset to ensure this claim.

A substantial limitation of the study is that the dataset was split into a train, test and validation set on slice-level instead of patient-level. This implies that slices from the same patient may have occurred in different subsets. Consequently, the experiments have potentially suffered from data leakage; since slices of the same patient can be strongly correlated, the model was partly tested on data that it had already seen - to some extent- in the training set. This could have introduced falsely elevated performance results. [107]

The dataset was divided into a high and low subtlety-score subset based on the mean subtlety score to test the robustness of the models. However, considerable variability in the radiologists' assessment of the subtlety score for the same nodules exists. [79]. Additionally, instead of categorizing each slice of a nodule, the subtlety score is only given for the entire nodule. Therefore, all slices of the same nodule obtained the same mean subtlety score. This does not always hold true, especially because nodules in slices further away from the center are often much smaller than in the middle slice. Consequently, the results in Figure 4.6 may have given a distorted perception of the reality. Instead of dividing the dataset into high and low subtlety-score to test the models' robustness, it could also be interesting to divide the dataset into each type of nodule separately based on its surrounding.(i.e. well-circumscribed, juxta-vascular, juxta-pleural) or its internal texture (GGO, part-solid, solid). This would provide additional insights about the network's performance on difficult types of nodules. Unfortunately, only the internal texture of the nodules was provided by the radiologists, not their adhesion type. Therefore, a division based on subtlety score was chosen to include both a nodule's texture and surroundings to some extent.
The low subtlety-score nodules were highly underrepresented in the training dataset, as was illustrated in Figure 3.3. Training on such unbalanced dataset could cause difficulties for the model to capture strong feature representations for the segmentation of low subtlety-score nodules [39]. This could have contributed to the lower performance on these nodules for both models. Imaginably, GANs could have a valuable role in tackling this imbalanced dataset problem. As an example, Qin et al. [39] adopted adversarial networks to promote the samples' diversity. They employed a cGAN to produce patches of synthetic lung nodule CT-images based on semantic layouts for a more balanced dataset.

Although the database applied in this experiment is widely used, final subsets of the data can vary greatly between studies due to different pre-processing techniques. Some studies only use the middle slice of each nodule, while others use multiple slices. Due to all variances in subsets, it is difficult to provide a completely fair comparison between different studies. In this study, it was chosen to select all slices of a nodule that were still larger than three millimetres in diameter and were annotated by at least three radiologists, to obtain the largest possible training dataset. Since the nodule is an irregularly shaped sphere, the area near the boundaries of the nodule may not include any nodular tissue. [81] False positives may be produced when using these slices, leading to contamination of the training sets. Liu et al. [81] proposed to use only the three middle slices of a nodule to prevent this from happening. However, this would have drastically reduced the amount of training samples.

A known limitation of segmentation networks is their inadequacy of precision at the boundaries. This lack of precision is caused not only by the type of loss function, but also by imprecise GTs. [108] The GTs in this study are obtained from the annotations made by four radiologists. However, this annotation process is based on each radiologist's subjective judgement, leading to a great amount of inter-personal variability in the annotation process. [79] As the GTs are based on the annotations of all four radiologists, this variability highly effects the reference GT. Hence, the models' performance may be affected by such variation. Forcing the model to strictly learn from an imprecise GT could bring to over-learning or overfitting. [108] Possible solutions to tackle this problem are to develop a loss function with tolerance on the border pixels of the nodule or to use softmasks. Pezzano et al. [108] proposed to use a loss function that does not just calculate the loss element-wise, but calculates the loss between one pixel of the prediction and all the corresponding surrounding pixels in the GT. Wang et al. [109] employed a soft labeling technique to reduce the effect of lung nodule annotation uncertainty by the radiologists. They created soft masks by setting the difference between the masks annotated by the radiologists as an uncertainty area, instead of taking the 50% consensus contour, and obtained valuable results on the LIDC-IDRI dataset.

In this study, only the slices that contained nodules were extracted from a CT-scan during pre-processing to assess the networks' ability to segment lung nodules. However, from a clinical point of view, it would have been desirable to develop a system that is able to take an entire CT-scan as input, from which it automatically detects and segments the slices that incorporate nodules. Since a CT-scan consists of slices with and without nodules, images that do not contain any nodule at all should have been added to the dataset in order for the models to learn the desired output on these images (completely black) as well. Performance evaluation on these slices too is necessary to assess the true detection ability of the models.

6 Conclusions and Future work

Image generation using GANs is currently an active research area. Although GANs are most commonly used to generate images with continuous values [47], several researchers explored their use for creating segmentation predictions with discrete labels. [44, 48, 70] The objective of this thesis was to implement a GAN-based segmentation method to automatically segment lung nodules from CT-images and to investigate to what extent this method could be of added value for improved segmentation performance compared to existing techniques. In order to alleviate the challenges arising from nodule segmentation and GAN training, a GAN architecture with an additional Dice loss, ASPP modules and a multiscale-L1 critic was proposed. The thesis concludes by answering the three sub-questions and the main research question proposed in the Introduction.

The first sub-question to be answered is: *"To what extent do the additional modules solve the segmentation challenges and contribute to a more accurate segmentation performance of the proposed method?"* Based on the ablation study, it can be concluded that a Dice loss improves the performance of the segmentation models and is of essential value in the class imbalance problem. The addition of ASPP modules to the generator network did not not contribute to an improved performance, which suggests that ASPP modules are not advantageous for lung nodule segmentation from entire CT-images. Lastly, the findings confirm that a multiscale-L1 critic was able to stabilize GAN training, in contrast to a conventional GAN.

The second sub-question that was researched was: *"How does the performance of the proposed method compare to the other GAN-based methods for lung nodule segmentation?"* The present findings show that the proposed GAN outperforms the CTumorGAN for lung nodule segmentation on entire image slices, although this conclusion is based on different datasets. Nevertheless, our GAN does not surpass the performance of patch-based GANs.

The last sub-question was: *"What are the performance differences between the proposed method and the U-Net and could the proposed method be of added value for the task of automatic lung nodule segmentation compared to this state-of-the-art segmentation network?"* This aspect of the research suggests that, although we can observe from other studies that GANs may have a valuable potential for medical image segmentation [44, 70, 110], the GAN developed in this research did not show superior performance to state-of-the-art U-Net segmentation models. Accordingly, we have to conclude that our proposed GAN is not of added value for the task of automatic lung nodule segmentation on entire CT-images. The U-Net implemented in this study achieved outstanding performance and even exceeded existing segmentation methods on Dice score.

These sub-questions helped to answer the main research question of this thesis: *To what extent can a GAN-based network be used for automatic segmentation of lung nodules in entire 2D CT-images?* To conclude the overall thesis and answer the main question; a multiscale-L1 GAN with an additional Dice loss was successfully implemented and used to automatically segment lung nodules from entire 2D CT-images. Although this GAN showed sufficient performance and comparable results to other GAN-based methods, it did not outperform the proposed U-Net model on our specific task. Still, the research into GANs is exciting and flourishing, and it is believed that, in the future, GANs could have a great potential to become an outstanding method for other medical segmentation purposes.

6.1 Future work

Some recommendations for future work can be made to further improve the proposed models as well as deep learning models in general in this area of study. These recommendations could advance the research and development of accurate automatic lung nodule segmentation methods.

A logical next step of this study is to extend the 2D U-Net to a 3D U-Net. During reading, radiologists visualize the CT-scans slice by slice, taking into account the 3D information. [108] It would be desirable if a neural network could to the same. Instead of analyzing each slice separately, a 3D model analyzes the input as volumetric and utilizes global features between the input slices. Therefore, a 3D model might extract more and richer spatial information of the nodules than a 2D model. [111] However, 3D models have the major disadvantage of increased complexity, computational cost and memory storage. [111] In addition, the irregularity in slice thickness could influence the 3D nodule segmentation, whereas segmentation in 2D images is not affected by slice thickness. [112] It is crucial to consider these limitations in the development of an efficient 3D model.

Regarding the GAN in specific, future research could focus on exploring new multi-scale feature loss functions. It might prove important to implement more suitable loss functions, which are less influenced by the class-imbalance problem than the L1 loss, to achieve improved performance of the GAN itself.

Neural networks are trained based on the information they are provided with. Therefore, efforts should be made to develop a more trustable and richer database. Future work may include solving the imbalanced dataset problem. A more balanced dataset could advance the model to learn characteristic features of all possible types of nodules and improve its performance on the underrepresented, low subtlety-score nodules. By expanding and balancing the dataset, a more generalized and robust model would be achieved. Image generation by GANs may be a good alternative to face the imbalanced dataset problem. Furthermore, it is advisable to add slices that do not contain nodules to the dataset to enable the development of models that can accurately detect and segment lung nodules from entire CT-scans.
Since the created annotations in the LIDC-IDRI database are made manually, errors will always exist. Therefore, a final future outlook would be to develop robust methods that are able to work and learn from noisy, imprecise GTs, for example by using more tolerant loss functions or by applying soft masks instead of binary masks.

Considering all the above recommendations, it may be possible to develop a robust and accurate clinical tool for automated nodule detection and segmentation. In the future, it is expected that such a tool has the potential to be implemented in clinical practice to enhance the manual nodule annotations and to reduce the reading time of the radiologists when used as an assistance tool.

Appendices

A Theoretical solution GAN training

This section shows the derivation of the optimal discriminator D and the optimal generator G, and concludes with the theoretical solution to the minimax game.

A.1 Derivation optimal D

The minimax game in equation 2.1 can be reformulated as:

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

$$= \mathbb{E}_{x\sim p_{data}}[\log D()] + \mathbb{E}_{x\sim p_g}[\log(1 - D(x))] \tag{A.2}$$

$$= \int_x \Big(p_{data}(x)\log D(x) + p_g(x)\log(1 - D(x))\Big)\, dx \tag{A.3}$$

For any (a,b) $\in \mathbb{R}^2$, the function $y = a\ \log(y) + b\ \log(1 - y)$ achieves its maximum at $\frac{a}{a+b}$. [49] This can be proven by the following derivation. The optimal y^* can be found by taking the first derivative, y', and equalizing this to zero:

$$\begin{aligned} y &= a\ \log(y) + b\ \log(1 - y) \\ y' &= \frac{a}{y} + \frac{b}{1 - y} \\ \frac{a}{y^*} &= \frac{b}{1 - y^*} \\ \frac{b}{a} &= \frac{1 - y^*}{y^*} \\ \frac{1}{y^*} &= \frac{a + b}{a} \\ y^* &= \frac{a}{a + b} \end{aligned}$$

By applying this to Equation A.3, the optimal discriminator $D_G(x)$ for fixed G can be derived:

$$D^*_G(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \tag{A.4}$$

A.2 Derivation optimal G

The optimal G, G^*, can be calculated by minimizing Equation 2.1 with respect to G and implementing the optimal discriminator:

$$\begin{aligned} G^* &= \min_G V(D^*_G, G) \\ &= \int_x \Big(p_{data}(x)\log(\frac{p_{data}(x)}{p_{data}(x) + p_g(x)}) + p_g(x)\log(\frac{p_g(x)}{p_{data}(x) + p_g(x)})\Big)\, dx \end{aligned}$$

This problem can be solved using the Jensen-Shannon-Divergence, which is defined as [72]:

$$D_{JS}(p_{data} \parallel p_g) = \frac{1}{2}D_{KL}(p_{data} \parallel \frac{p_{data} + p_g}{2}) + \frac{1}{2}D_{KL}(p_g \parallel \frac{p_{data} + p_g}{2}) \tag{A.5}$$

Where D_{KL} is the Kullback-Leibler divergence [72]:

$$D_{KL}(p \parallel q) = \int p(x)\ \log(\frac{p(x)}{q(x)})\ dx \tag{A.6}$$

Rewriting equation A.5 using the definition of the Kullback-Leibler divergence results in:

$$
\begin{aligned}
D_{JS}(p_{data} \parallel p_g) &= \frac{1}{2}\Big(\int_x p_{data}(x)\ \log(\frac{2p_{data}(x)}{p_{data}(x)+p_g(x)})\ dx\Big) + \frac{1}{2}\Big(\int_x p_g(x)\ \log(\frac{2p_g(x)}{p_{data}(x)+p_g(x)})\ dx\Big) \\
&= \frac{1}{2}\Big(\int_x p_{data}(x)\ (\log\ 2 + \log(\frac{p_{data}(x)}{p_{data}(x)+p_g(x)}))\ dx\Big) + \\
&\quad \frac{1}{2}\Big(\int_x p_g(x)\ (\log\ 2 + \log(\frac{2p_g(x)}{p_{data}(x)+p_g(x)}0)\ dx\Big) \\
&= \frac{1}{2}\Big(\log\ 2 + \int_x p_{data}(x)\ \log(\frac{p_{data}(x)}{p_{data}(x)+p_g(x)})\ dx\Big) + \\
&\quad \frac{1}{2}\Big(\log\ 2 + \int_x p_{data}(x)\ \log(\frac{p_g(x)}{p_{data}(x)+p_g(x)})\ dx\Big) \\
&= \frac{1}{2}\Big(\log\ 4 + \min_G V(D^*_G, G)\Big)
\end{aligned}
$$

Thus:

$$
\min_G V(D^*_G, G)) = -\log\ 4 + 2D_{JS}(p_{data} \parallel p_g) \tag{A.7}
$$

As the Jensen-Shannon divergence between two distributions is always positive and zero only when the distributions are equal [72], $G^* = -\log 4$ is the global minimum of the equation above and thus the solution for the optimal generator. This is achieved if and only if $p_g = p_{data}$. The optimal values for D and V then are:

$$
\begin{aligned}
D^*_G(x) &= \frac{p_{data}}{p_{data}+p_g} = \frac{1}{2} \\
\min_G \max_D V(D,G) &= \mathbb{E}_{x\sim p_{data}}[\log\ \frac{1}{2}] + \mathbb{E}_{\mathbf{z}\sim p_z}[\log\ (1-\frac{1}{2}] \\
&= -2\ \log\ 2
\end{aligned}
$$

B Additional results ablation study

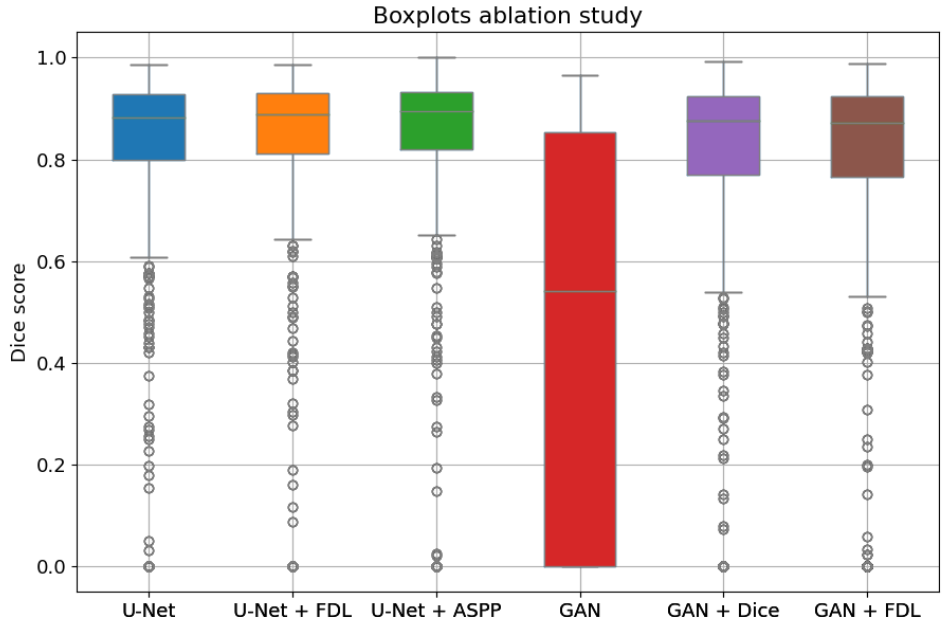


Figure B.1: Boxplots showing the distribution of dice scores for the various ablation study methods.

B.1 Loss functions

The performance results of all the loss functions tested during the ablation study are given in Table 7. The main differences occur in the precision and recall values. This is visualized in a Precision-Recall plot in Figure B.2. The precision and recall were determined at 25 binarization thresholds and plotted against each other. For binary segmentation, the iso-f1 curves correspond to the dice score. Figure B.2 shows that the most substantial differences occur between the Tversky losses. This is as expected, as the loss functions both focus on opposite predictions. In agreement with the results found in Table 7, all loss function curves lie around the same f1 curve, although the theoretical f1 values are slightly larger than the dice values found during the experiments.

Table 7: Quantitative performance results of the U-Net with various loss functions.

Loss	DSC	Precision	Recall
Dice	0.8108	0.7984	0.8462
FDL ($\gamma = 0.5$)	**0.8143**	0.8008	0.8493
FDL ($\gamma = 2.0$)	0.8062	**0.8056**	0.8329
Tversky ($\alpha = 0.7$, $\beta = 0.3$)	0.8069	0.8415	0.7942
Tversky ($\alpha = 0.3$, $\beta = 0.7$)	0.8051	0.7680	0.8755
FTL ($\gamma = 0.5$, $\alpha = 0.3$, $\beta = 0.7$)	0.8084	0.7741	0.8676
FTL ($\gamma = 2.0$, $\alpha = 0.3$, $\beta = 0.7$)	0.8051	0.7680	**0.8755**

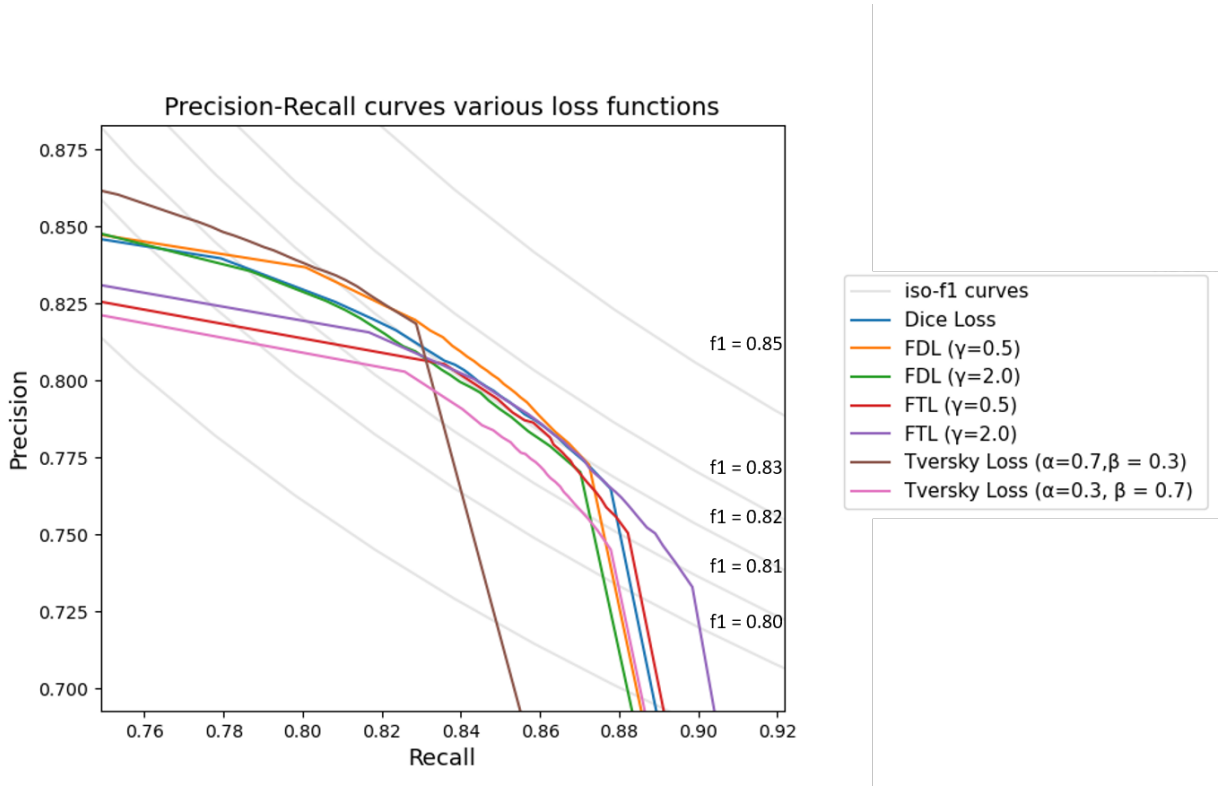


Figure B.2: Precision-Recall curves for various loss functions. Precision and recall calculated at different binarization thresholds.

B.2 ASPP Gaussian

Dilated convolutions might be thought of as the combination of a sampling procedure and a common convolution. From this point of view, it could be beneficial to introduce Gaussian convolutions in the ASPP module. In this study, ASPP modules containing Gaussian convolutions were implemented in the U-Net and compared to the results of the ASPP modules without Gaussian convolutions. For each dilated convolution, a Gaussian convolution G with standard variance σ and truncation t was performed on the feature map following the dilated convolution. Therefore, the formula for dilated convolution (Equation 3.9) extended to:

$$\mathbf{y}[\mathbf{i}] = \sum_{(\mathbf{k}} G((\mathbf{x}[\mathbf{i} + r \cdot \mathbf{k}]\mathbf{w}[\mathbf{k}]), \sigma, t)$$

The results are shown in Figure B.3. For the ASPP modules in the downsampling path, the ASPP Gaussian showed worse performance on all evaluation metrics than the standard ASPP. For the other two locations, the Gaussian ASPP showed a slightly better performance on the dice and precision, but not on its recall score. The Gaussian ASPP thus does not consistently perform better than the basic ASPP. As the image size is not scaled down during an ASPP procedure and the resolution remains equal, dilated convolutions are generally not viewed as a sampling procedure. From the results found in this study, it can not be concluded that applying a Gaussian convolution after dilated convolutions is essential. It is expected that the performance increase (occuring especially in the skipconnection) could mainly be explained by the fact that nodules are extremely small and could benefit from a more smoothed representation, as this enlarges the tumor regions. This potential benefit was also shown by Shen et al. [105]. They used Gaussian kernels preceding dilated convolutions to produce more robust feature representations. The Gaussian kernel was performed on the feature maps to accumulate information in each position and capture context.

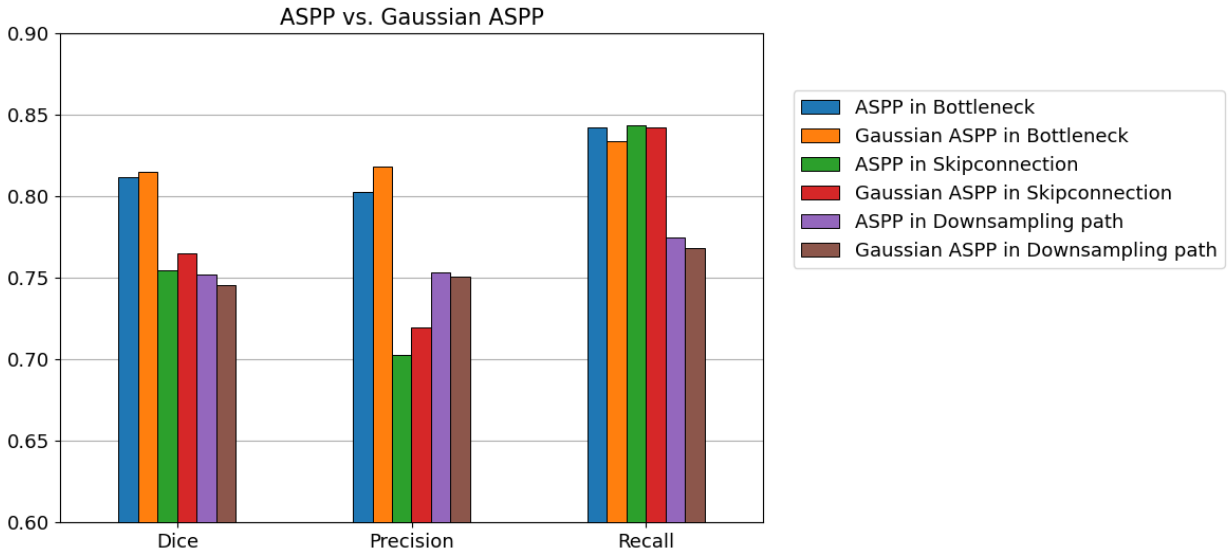


Figure B.3: Performance comparison of ASPP modules and Gaussian ASPP modules, implemented at various locations.

C Results GAN without additional dice loss

The GAN without an additional dice loss, using only adversarial training based on the multiscale L1 loss, did not show acceptable results. A histogram of all dice values produced by this GAN is demonstrated in Figure C.1. It is noticeably that the GAN's poor performance originates due to many non-overlapping predictions, resulting in a DSC value of 0. These were mainly images in which the model's output was completely black, i.e. the model did not predict any tumor pixels at all. The instances in which the model did predict tumor pixels, the predicted output was quite adequate. This is also visualized in Figure C.2, in which some examples of the model's output compared to the GT are shown. For the two images on the left, the model did not predict any tumor pixels, while the images on the right show good correspondence to the GT.

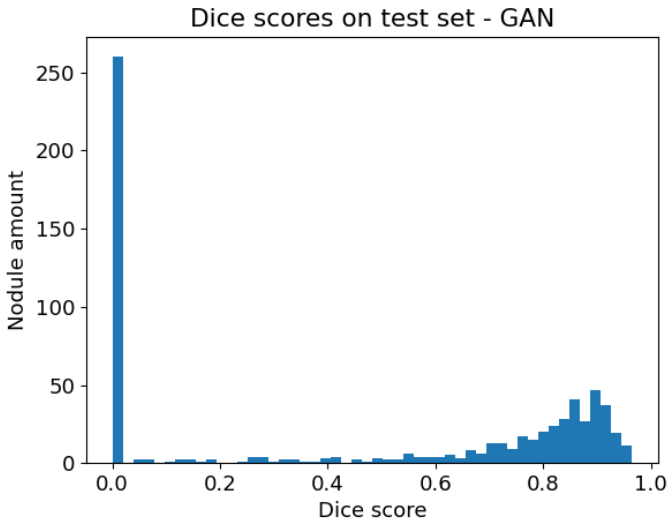


Figure C.1: Histogram of dice values

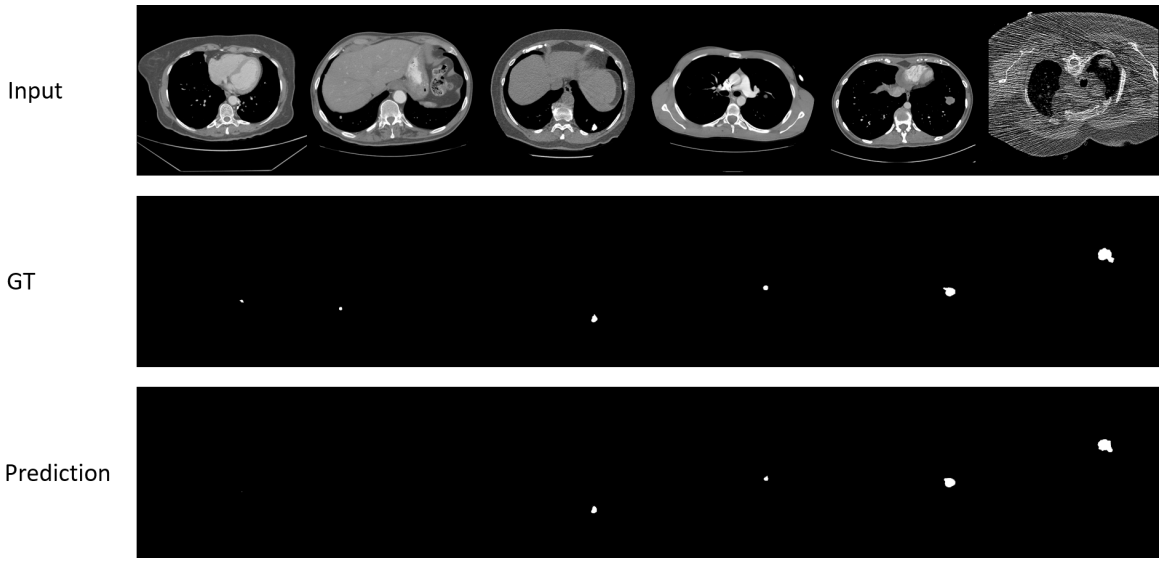
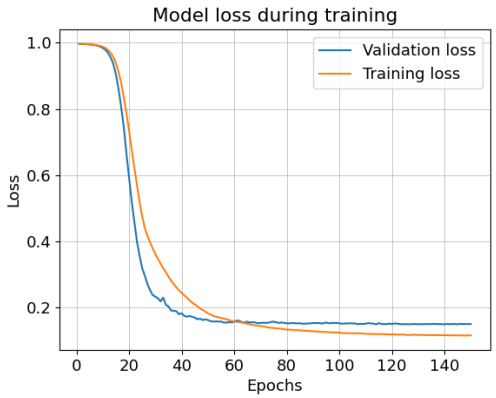


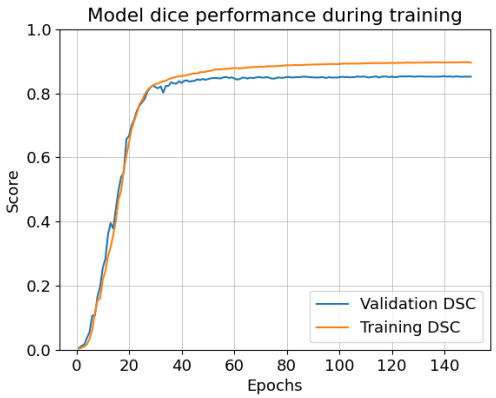
Figure C.2: Visual performance results of the GAN. Top row shows the input CT-images, middle row the GT masks and bottom row the segmentation maps predicted by the model.

D Training graphs

The training progress of the U-Net, GAN and GAN + Dice are shown in Figure D.1, Figure D.2, and Figure D.3, respectively. The losses and dice scores are plotted as a function of training epochs.

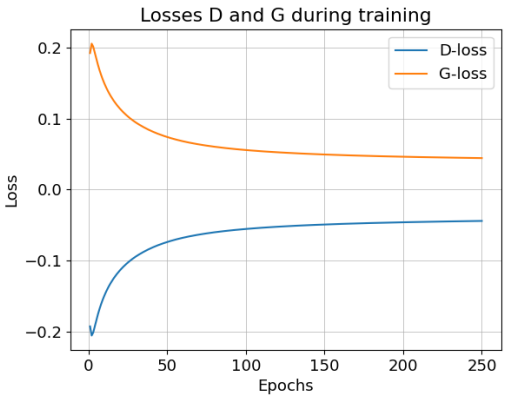


(a) Progress of loss for train and validation set.

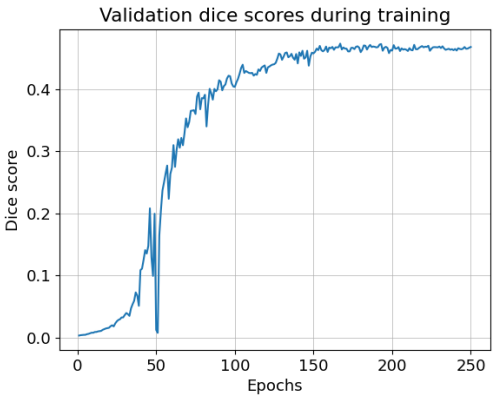


(b) Dice score on train and validation set during training.

Figure D.1: Training progress of the U-Net

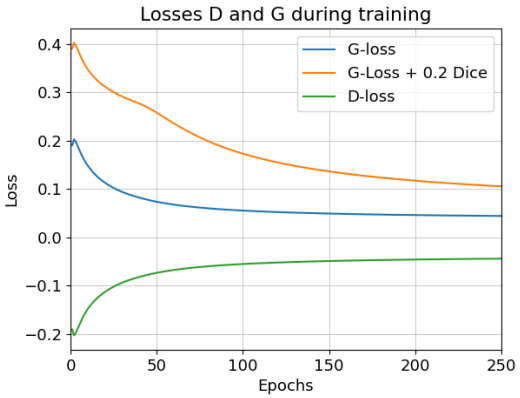


(a) Progress of generator and discriminator loss.

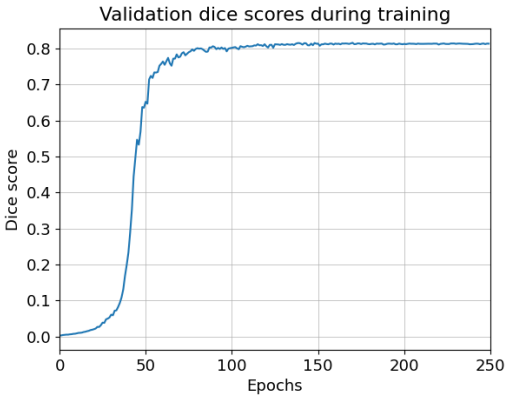


(b) Dice score on validation set during training.

Figure D.2: Training progress of the GAN



(a) Progress of generator and discriminator loss.



(b) Dice score on validation set during training.

Figure D.3: Training progress of the GAN with additional dice loss

E Comparison patch-based models

Table 8: Comparison of the propsed U-Net with state-of-the-art patch-based models. The results are in the format of "mean dice ± standard deviation". The proposed U-Net is indicated in bold.

Method	Year	Patch-size	DSC (%)
2D modified residual U-Net [42]	2019	64×64	70.9
Multi-task learning CNN [41]	2018	64×64	73.89
Multi-view deep CNN [28]	2017	35×35	77.67 ± 15.71
Cascaded dual-pathway residual network [113]	2019	35×35	81.58 ± 11.05
Central focussed CNN [114]	2017	35×35	82.15 ± 10.76
Context-learning (CoLe) CNN [108]	2021	64×64	82.5
Dual-branch residual network [29]	2019	35×35	82.74 ± 10.19
NoduleNet [115]	2019	64×64	83.10 ± 8.85
Proposed U-Net	**2021**	$\mathbf{512 \times 512}$	$\mathbf{84.33 \pm 16.44}$
3D-CNN + cGAN for data augmentation [39]	2019	64×64	84.83
Adaptive ROI with multi-view residual learning [116]	2020	128×128	87.55 ± 10.58
Deep Learned Shape Driven Level Set[1] [112]	2019	128×128	93.00 ± 11.00

[1] Applied on solid nodules only

F Multiscale L2 GAN

Multiple experiments were performed with the GAN-based model in which the multiscale L1 loss was replaced by a multiscale L2 loss. The ℓ_{mae} in Equation 3.2 was replaced by a Mean Squared Error (MSE), also known as the L_2 distance:

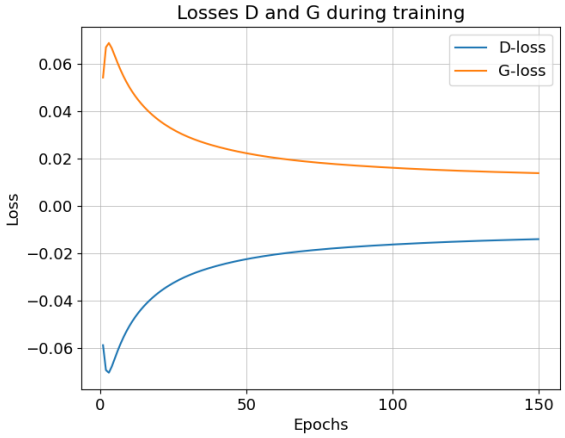
$$\ell_{mse}(f_C(x), f_C(\hat{x}) = \frac{1}{L}\sum_{i=1}^{L}(f_C^i(x) - f_C^i(\hat{x}))^2 \tag{F.1}$$

where L represents the total number of layers (scales) in the critic, and $f_C^i(x)$ the extracted feature map of image x at layer i.

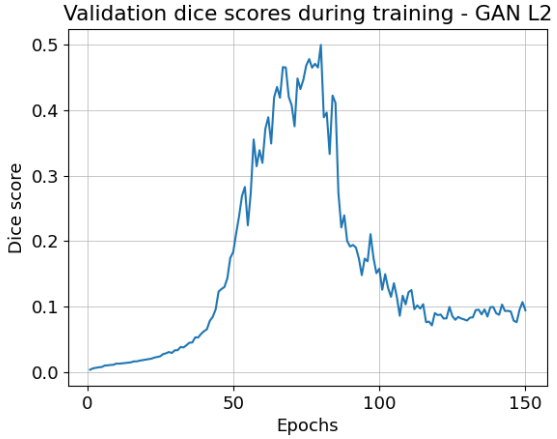
The results of this multiscale L2 GAN were compared to the multiscale L1 GAN, as can be seen in Table 9. In combination with an additional dice loss, the multiscale L1 loss showed best performance. The dice loss was scaled to 0.2 and 0.06 to be in balance with the magnitude of the multiscale L1 and L2 loss, respectively. Without additional dice loss, the L2 loss outperformed the L1 loss. However, the training progress, visualized in Figure F.1, did not show a stable course. After about 75 epochs, the validation dice dropped and the model was not able to learn any further. We also experimented with a larger batch size (BS = 20), a smaller batch size (BS = 6), disabling parameter clipping and disabling dropout, but none of these helped to stabilize training. Therefore, the L1 loss was chosen to be used in this thesis.

Table 9: Performance evaluation of multiscale L1 and multiscale L2 loss. Values are based on two runs and are given as mean ± std.

Loss	DSC	Precision	Recall
L1	0.4707 ± 0.0008	**0.5136 ± 0.0023**	0.4705 ± 0.0048
L2	**0.5053 ± 0.0207**	0.4742 ± 0.0156	**0.5914 ± 0.0290**
L1+0.2Dice	**0.8110 ± 0.0040**	**0.8240 ± 0.0045**	0.8212 ± 0.0048
L2+0.06Dice	0.8008 ± 0.0035	0.8047 ± 0.0012	**0.8265 ± 0.0031**



(a) Progress of generator and discriminator loss.



(b) Dice score on validation set during training.

Figure F.1: Training progress of the Multiscale L2 GAN

G Unstable GANs

In addition to the multiscale L1 critic, some more conventional GANs were also implemented for the task of lung nodule segmentation in this study: a vanilla GAN, a LSGAN and a WGAN. Furthermore, the openly-available Pix2Pix algorithm was applied for our task. Some examples of the training courses of these GANs are shown in this section. All of these GANs failed to produce stable results. Parameter tuning was done by adapting the learning rate, batch size, dropout ratio, parameter clipping etc., but unfortunately none of this helped.

Vanilla GAN The vanilla GAN was implemented with a BCE-Loss + L1-Loss and additionally label smoothing was applied. It achieved a best mean dice score of 0.0439. The discriminator and generator loss rapidly became constant, signifying that they do not improve each other.

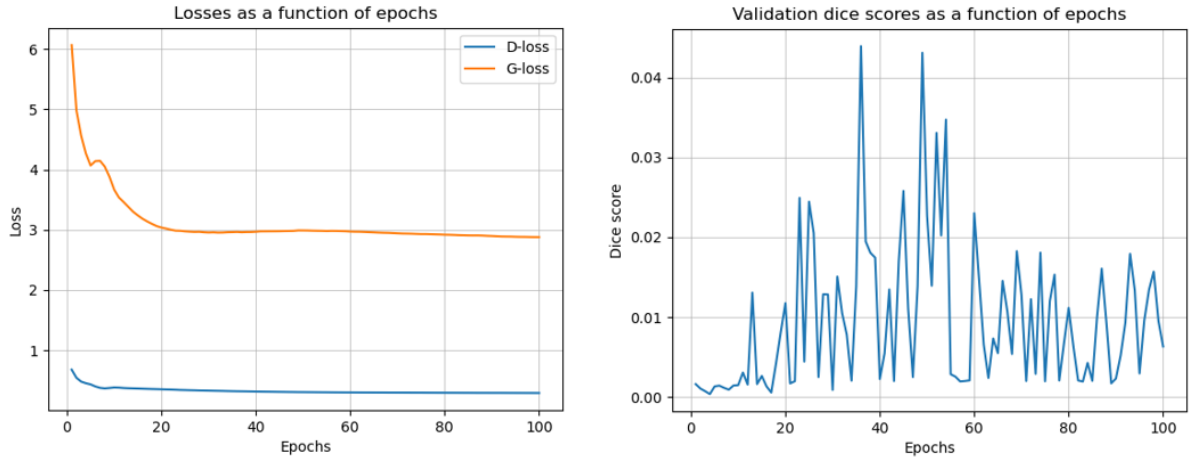


Figure G.1: Training process vanilla GAN.

LSGAN The LSGAN obtained a best mean dice score of 0.0125. The GAN failed to converge, as the discriminator loss quickly dropped to zero, while the generator loss kept rising. This is probably caused by the generator outputting lousy images that the discriminator can easily identify as fake.

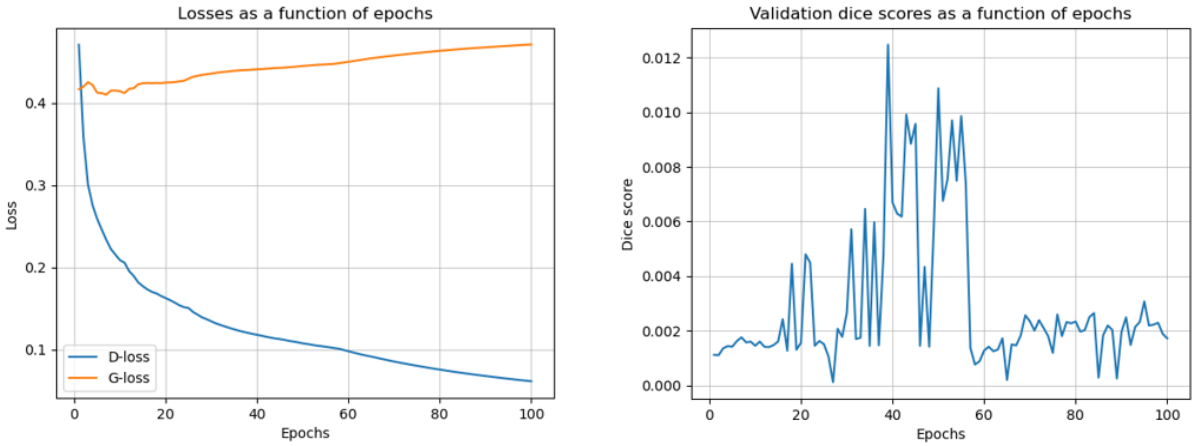


Figure G.2: Training process LSGAN.

WGAN At last, the WGAN also showed an unstable course and obtained a best mean dice score of 0.0044.

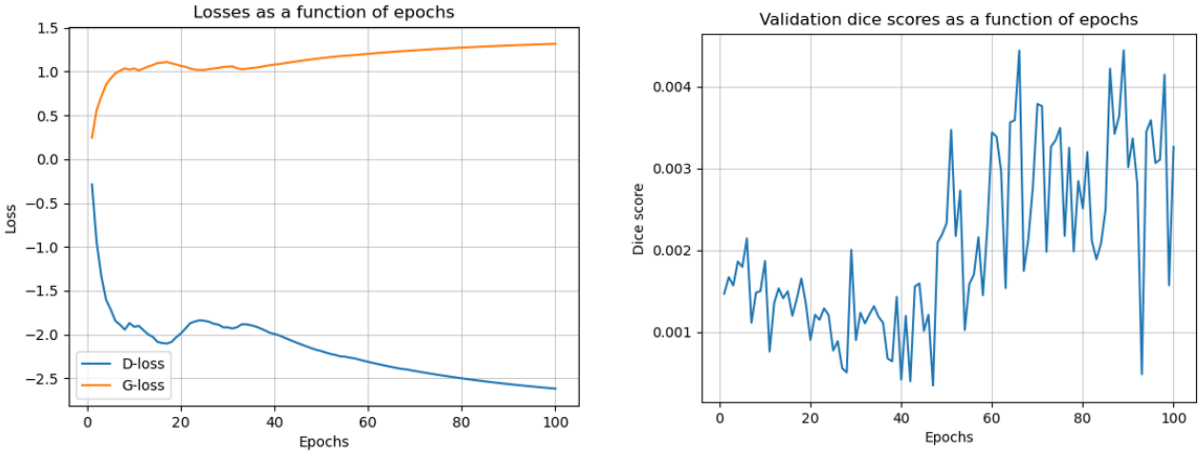


Figure G.3: Training process WGAN.

G.1 Pix2Pix

The Pix2Pix framework is openly available from [117] and adapted for our dataset. Three options of the Pix2Pix framework were tested: Vanilla GAN, LSGAN, WGAN. The adversarial loss was combined with an L1 loss, as was used by Isola et al. [47].

Pix2Pix - Vanilla The vanilla Pix2Pix GAN obtained a best average dice score of 0.0030. Its minimum dice was 0.0, while its maximum dice was 0.4390. As can be seen from Figure G.4, the discriminator loss quickly drops to zero, while the generator loss increases over the iterations. The training does not converge.

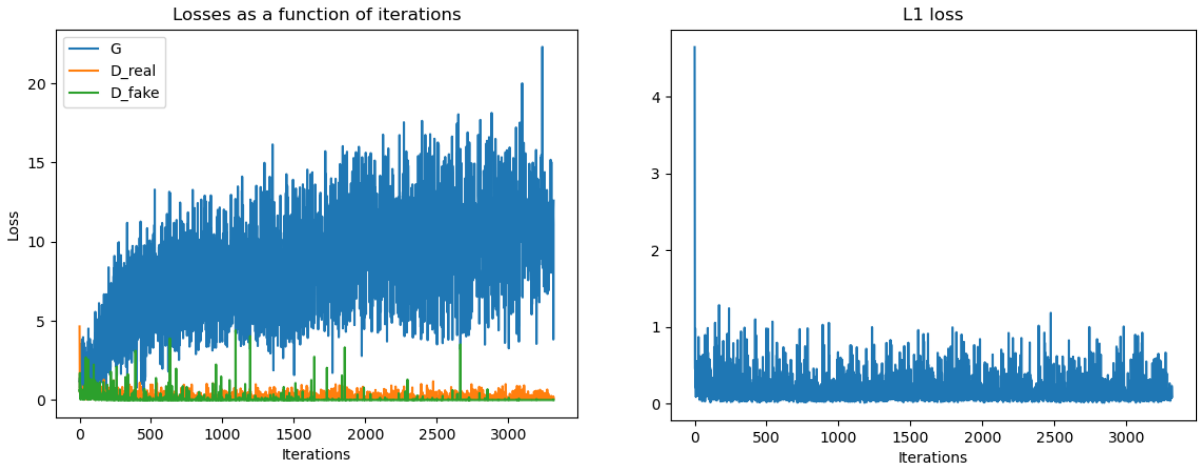


Figure G.4: Pix2Pix- Vanilla GAN.

Pix2Pix - LSGAN The LSGAN showed mode collapse; Almost all predicted outcomes were visually the same. The mode collapse is also evident in the training graphs. The losses suddenly drop after about 3500 iterations. Before the mode collapse, the average dice score was 0.0018, with a maximum value of 0.4694.

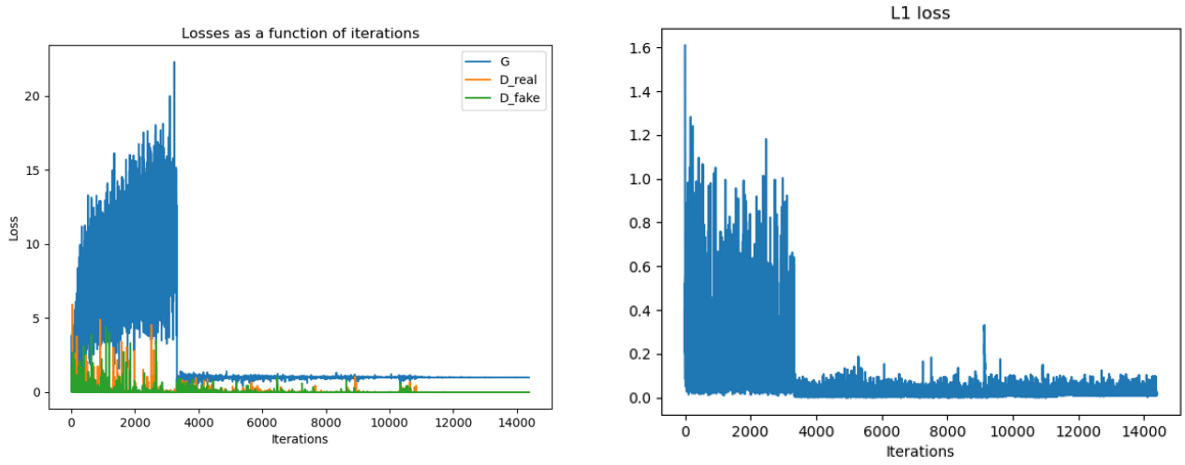


Figure G.5: Pix2Pix- LSGAN.

Pix2Pix - WGAN The WGAN Pix2Pix model showed an incomprehensible training course. It seemed to do nothing for about 15000 iterations, after which training suddenly arose. This is mainly the case because previous loss function values are negligible compared to the extremely large values after 15000 iterations. The mean dice score of this WGAN was 0.0028, with a maximum score of 0.6246.

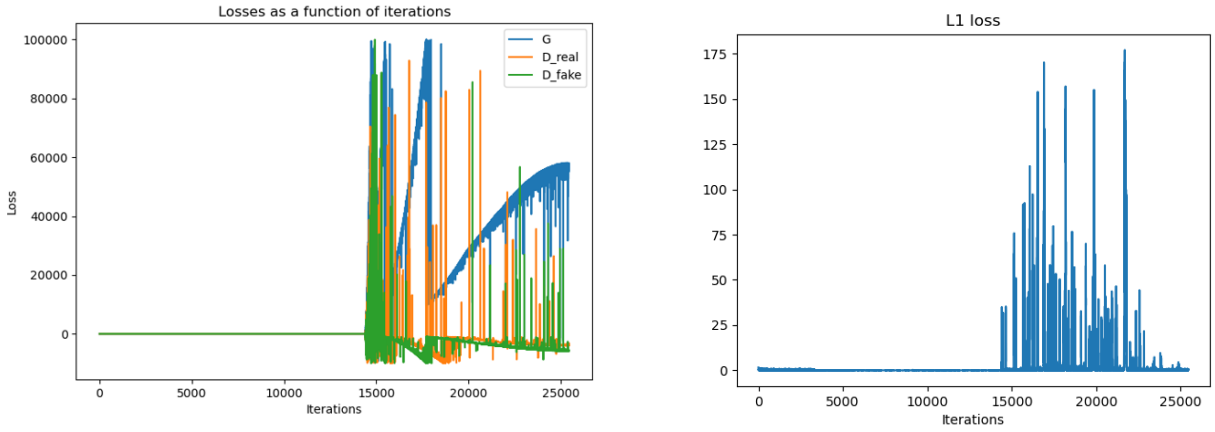


Figure G.6: Pix2Pix- WGAN.

References

[1] World Health Organization. Cancer; 2018. Available from: https://www.who.int/news-room/fact-sheets/detail/cancer.

[2] Meraj T, Rauf HT, Zahoor S, Hassan A, Lali MIU, Ali L, et al. Lung nodules detection using semantic segmentation and classification with optimal features. Neural Computing and Applications. 2020 may:1–14. Available from: https://doi.org/10.1007/s00521-020-04870-2.

[3] Midthun DE. Early detection of lung cancer. F1000Research. 2016;5. Available from: /pmc/articles/PMC4847569/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4847569/.

[4] Makaju S, Prasad PWC, Alsadoon A, Singh AK, Elchouemi A. Lung Cancer Detection using CT Scan Images. Procedia Computer Science. 2018;125(2009):107–114. Available from: https://doi.org/10.1016/j.procs.2017.12.016.

[5] Loverdos K, Fotiadis A, Kontogianni C, Iliopoulou M, Gaga M. Lung nodules: A comprehensive review on current approach and management. Annals of Thoracic Medicine. 2019 10;14:226.

[6] El-Baz A, Beache GM, Gimel'Farb G, Suzuki K, Okada K, Elnakib A, et al. Computer-aided diagnosis systems for lung cancer: Challenges and methodologies. International Journal of Biomedical Imaging. 2013;2013.

[7] Shi Z, Hu Q, Yue Y, Wang Z, AL-Othmani OMS, Li H. Automatic Nodule Segmentation Method for CT Images Using Aggregation-U-Net Generative Adversarial Networks. Sensing and Imaging. 2020;21(1):1–16.

[8] Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Scientific Reports. 2019 dec;9(1):1–8. Available from: https://doi.org/10.1038/s41598-019-40041-7.

[9] Leader JK, Warfel TE, Fuhrman CR, Golla SK, Weissfeld JL, Avila RS, et al. Pulmonary Nodule Detection with Low-Dose CT of the Lung: Agreement Among Radiologists. AJR. 2005:185. Available from: www.ajronline.org.

[10] Koenigkam Santos M, Raniery Ferreira Júnior J, Tadao Wada D, Priscilla Magalhães Tenório A, Henrique Nogueira Barbosa M, Mazzoncini De Azevedo Marques P. Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: Advances in imaging towards to precision medicine. Radiologia Brasileira. 2019 nov;52(6):387–396. Available from: /pmc/articles/PMC7007049/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7007049/.

[11] Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, et al. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Scientific Reports. 2016;6(March):1–13. Available from: http://dx.doi.org/10.1038/srep24454.

[12] Razzak MI, Naz S, Zaib A. Deep learning for medical image processing: Overview, challenges and the future. Lecture Notes in Computational Vision and Biomechanics. 2018;26:323–350.

[13] Wolterink JM, Kamnitsas K, Ledig C, Išgum I. Deep learning: Generative adversarial networks and adversarial methods. In: Handbook of Medical Image Computing and Computer Assisted Intervention. Elsevier Inc.; 2019. p. 547–574. Available from: https://doi.org/10.1016/B978-0-12-816176-0.00028-4.

[14] Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Elsevier GmbH; 2019.

[15] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444.

[16] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems. 2012;25:1097–1105.

[17] LeCun Y, Haffner P, Bottou L, Bengio Y. Object recognition with gradient-based learning. Shape, contour and grouping in computer vision. 1999:319–345.

[18] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017;42(December 2012):60–88.

[19] Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annual Review of Biomedical Engineering. 2017 jun;19:221–248. Available from: /pmc/articles/PMC5479722/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5479722/.

[20] Wiemker R, Zwartkruis A. Optimal thresholding for 3D segmentation of pulmonary nodules in high resolution CT. International Congress Series. 2001 jun;1230(C):653–658.

[21] Kostis WJ, Reeves AP, Yankelevitz DF, Henschke CI. Three-Dimensional Segmentation and Growth-Rate Estimation of Small Pulmonary Nodules in Helical CT Images. IEEE Transactions on Medical Imaging. 2003 oct;22(10):1259–1274.

[22] Dehmeshki J, Amin H, Valdivieso M, Ye X. Segmentation of pulmonary nodules in thoracic CT scans: A region growing approach. IEEE Transactions on Medical Imaging. 2008 apr;27(4):467–480.

[23] Wang Q, Song E, Jin R, Han P, Wang X, Zhou Y, et al. Segmentation of Lung Nodules in Computed Tomography Images Using Dynamic Programming and Multidirection Fusion Techniques 1. Academic radiology. 2009 may;16:678–688. Available from: http://imaging.cancer.gov/.

[24] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;39(12):2481–2495.

[25] Lee SLA, Kouzani AZ, Hu EJ. Automated detection of lung nodules in computed tomography images: A review. Machine Vision and Applications. 2012;23(1):151–163.

[26] Li Y, Shen L. CC-GAN: A Robust Transfer-Learning Framework for HEp-2 Specimen Image Segmentation. IEEE Access. 2018;6:14048–14058.

[27] Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Deep neural networks segment neuronal membranes in electron microscopy images. Advances in Neural Information Processing Systems. 2012;4:2843–2851.

[28] Wang S, Zhou M, Gevaert O, Tang Z, Dong D, Liu Z, et al. A multi-view deep convolutional neural networks for lung nodule segmentation. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. 2017:1752–1755.

[29] Cao H, Liu H, Song E, Hung CC, Ma G, Xu X, et al. Dual-branch residual network for lung nodule segmentation. Applied Soft Computing Journal. 2020 jan;86:105934.

[30] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing; 2015. p. 234–241.

[31] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015;39(4):640–651.

[32] Hesamian MH, Jia W, He X, Kennedy P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. Journal of Digital Imaging. 2019;32(4):582–596.

[33] Yang D, Xiong T, Xu D, Kevin Zhou S. Segmentation using adversarial image-to-image networks. Elsevier Inc.; 2019. Available from: https://doi.org/10.1016/B978-0-12-816176-0.00012-0.

[34] Anthimopoulos M, Christodoulidis S, Ebner L, Geiser T, Christe A, Mougiakakou S. Semantic Segmentation of Pathological Lung Tissue With Dilated Fully Convolutional Networks. IEEE Journal of Biomedical and Health Informatics. 2019;23(2):714–722.

[35] Hossain S, Najeeb S, Shahriyar A, Abdullah ZR, Ariful Haque M. A Pipeline for Lung Tumor Detection and Segmentation from CT Scans Using Dilated Convolutional Neural Networks. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 2019;2019-May:1348–1352.

[36] Jadon S. A survey of loss functions for semantic segmentation. 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2020. 2020 jun. Available from: http://arxiv.org/abs/2006.14822http://dx.doi.org/10.1109/CIBCB48159.2020.9277638.

[37] Lan T, Li Y, Murugi JK, Ding Y, Qin Z. RUN:Residual U-Net for Computer-Aided Detection of Pulmonary Nodules without Candidate Selection. arXiv. 2018 may. Available from: http://arxiv.org/abs/1805.11856.

[38] Keetha NV, Anosh Babu S, Sekhara C, Annavarapu R. U-DET: A modified U-Net architecture with bidirectional feature network for lung nodule segmentation; 2020.

[39] Qin Y, Zheng H, Huang X, Yang J, Zhu YM. Pulmonary nodule segmentation with CT sample synthesis using adversarial networks. Medical Physics. 2019;46(3):1218–1229.

[40] Pang S, Du A, He X, Díez J, Orgun MA. Fast and accurate lung tumor spotting and segmentation for boundary delineation on CT slices in a coarse-to-fine framework. Communications in Computer and Information Science. 2019;1142 CCIS:589–597.

[41] Wu B, Zhou Z, Wang J, Wang Y. Joint learning for pulmonary nodule segmentation, atributes and malignancy prediction. In: IEEE International Symposium on Biomedical Imaging; 2018. p. 1109–1113.

[42] Carvalho JBS, Moreira JM, Figueiredo MAT, Papanikolaou N. Automatic detection and segmentation of lung lesions using deep residual CNNs. Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019. 2019:977–983.

[43] Luc P, Couprie C, Chintala S, Verbeek J. Semantic Segmentation using Adversarial Networks. 2016. Available from: http://arxiv.org/abs/1611.08408.

[44] Xue Y, Xu T, Zhang H, Long LR, Huang X. SegAN: Adversarial Network with Multi-scale L 1 Loss for Medical Image Segmentation. Neuroinformatics. 2018;16(3-4):383–392.

[45] Kohl S, Bonekamp D, Schlemmer HP, Yaqubi K, Hohenfellner M, Hadaschik B, et al. Adversarial Networks for the Detection of Aggressive Prostate Cancer. 2017. Available from: http://arxiv.org/abs/1702.08014.

[46] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Medical Image Analysis. 2019;58.

[47] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. 2017;2017-Janua:5967–5976.

[48] Lei B, Xia Z, Jiang F, Jiang X, Ge Z, Xu Y, et al. Skin lesion segmentation via generative adversarial networks with dual discriminators. Medical Image Analysis. 2020;64.

[49] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in Neural Information Processing Systems. 2014;3(January):2672–2680.

[50] Chuquicusma MJM, Hussein S, Burt J, Bagci U. How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. Proceedings - International Symposium on Biomedical Imaging. 2018;2018-April:240–244.

[51] Beers A, Brown J, Chang K, Campbell JP, Ostmo S, Chiang MF, et al. High-resolution medical image synthesis using progressively grown generative adversarial networks. 2018. Available from: http://arxiv.org/abs/1805.03144.

[52] Wolterink J, Leiner T, Viergever M, Išgum I. Generative adversarial networks for noise reduction in low-dose CT. IEEE transactions on medical imaging. 2017 12;36(12):2536–2545.

[53] Wang C, Xu R, Xu S, Meng W, Xiao J, Peng Q, et al. Accurate 2D soft segmentation of medical image via SoftGAN network.

[54] Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I. Deep MR to CT Synthesis Using Unpaired Data. In: Tsaftaris SA, Gooya A, Frangi AF, Prince JL, editors. Simulation and Synthesis in Medical Imaging. Cham: Springer International Publishing; 2017. p. 14–23.

[55] Dai W, Dong N, Wang Z, Liang X, Zhang H, Xing EP. Scan: Structure correcting adversarial network for organ segmentation in chest x-rays. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2018;11045 LNCS:263–273.

[56] Hu B, Tang Y, Chang E, Fan Y, Lai M, Xu Y. Unsupervised Learning for Cell-Level Visual Representation in Histopathology Images With Generative Adversarial Networks. IEEE Journal of Biomedical and Health Informatics. 2017 11;PP.

[57] Baumgartner CF, Koch LM, Tezcan KC, Ang JX, Konukoglu E. Visual Feature Attribution Using Wasserstein GANs. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2018:8309–8319.

[58] Schlegl T, Seeböck P, Waldstein SM, Schmidt-Erfurth U, Langs G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017;10265 LNCS:146–147.

[59] Moeskops P, Veta M, Lafarge MW, Eppenhof KAJ, Pluim JPW. Adversarial training and dilated convolutions for brain MRI segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017 jul;10553 LNCS:56–64. Available from: http://arxiv.org/abs/1707.03195.

[60] Sun L, Wang J, Huang Y, Ding X, Greenspan H, Paisley J. An adversarial learning approach to medical image synthesis for lesion detection. IEEE Journal of Biomedical and Health Informatics. 2020;24(8):2303–2314.

[61] Pang S, Du A, Orgun MA, Yu Z, Wang Y, Wang Y, et al. CTumorGAN: a unified framework for automatic computed tomography tumor segmentation. European Journal of Nuclear Medicine and Molecular Imaging. 2020;47(10):2248–2268.

[62] Zappa C, Mousa SA. Non-small cell lung cancer: Current treatment and future advances. Translational Lung Cancer Research. 2016 jun;5(3):288–300. Available from: /pmc/articles/PMC4931124/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4931124/.

[63] Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. Radiology. 2008;246(3):697–722.

[64] Ciompi F, Chung K, Van Riel SJ, Arindra A, Setio A, Gerke PK, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning OPEN. Scientific Reports. 2017 apr. Available from: www.nature.com/scientificreports.

[65] Small H, Ventura J. Handling Unbalanced Data in Deep Image Segmentation; 2017. Available from: https://svds.com/learning-imbalanced-classes/.

[66] Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016. 2016:565–571.

[67] Goodfellow I. NIPS 2016 Tutorial: Generative Adversarial Networks. 2016. Available from: http://arxiv.org/abs/1701.00160.

[68] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. Advances in Neural Information Processing Systems. 2016;(June):2234–2242.

[69] Xu C, Xu L, Brahm G, Zhang H, Li S. MuTGAN: Simultaneous segmentation and quantification of myocardial infarction without contrast agents via joint adversarial learning. vol. 11071 LNCS. Springer International Publishing; 2018. Available from: http://dx.doi.org/10.1007/978-3-030-00934-2{_}59.

[70] Wei Z, Shi F, Song H, Ji W, Han G. Attentive boundary aware network for multi-scale skin lesion segmentation with adversarial training. Multimedia Tools and Applications. 2020;79(37-38):27115–27136.

[71] Tang Z, Liu X, Li Y, Yap PT, Shen D. Multi-Atlas Brain Parcellation Using Squeeze-and-Excitation Fully Convolutional Networks. IEEE Transactions on Image Processing. 2020;29(May):6864–6872.

[72] Nielsen F. On a generalization of the Jensen-Shannon divergence. Entropy. 2020;22(2):221.

[73] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. 2017. Available from: http://arxiv.org/abs/1701.07875.

[74] Szegedy C, Vanhoucke V, Ioffe S, Shlens J. Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016:2818–2826.

[75] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings. 2016:1–16.

[76] Mirza M, Osindero S. Conditional Generative Adversarial Nets. 2014:1–7. Available from: http://arxiv.org/abs/1411.1784.

[77] Mao X, Li Q, Xie H, Lau R, Zhen W, Smolley S. Least Squares Generative Adversarial Networks; 2017. p. 2813–2821.

[78] Nowozin S, Cseke B, Tomioka R. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization; 2016.

[79] Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, 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. 2011;38(2):915–931.

[80] pylidc — pylidc documentation;. Available from: https://pylidc.github.io/{#}tutorials.

[81] Liu M, Dong J, Dong X, Yu H, Qi L. Segmentation of Lung Nodule in CT Images Based on Mask R-CNN. 2018 9th International Conference on Awareness Science and Technology, iCAST 2018. 2018:95–100.

[82] DenOtter TD, Schubert J. Hounsfield Unit. In: Definitions. Qeios; 2020. Available from: https://www.ncbi.nlm.nih.gov/books/NBK547721/.

[83] A Segmentation Framework of Pulmonary Nodules in Lung CT Images. Journal of Digital Imaging. 2016;29(1):86–103.

[84] Adams JE, Mughal Z, Damilakis J, Offiah AC. Radiology. In: Pediatric Bone. Elsevier Inc.; 2012. p. 277–307.

[85] Rahman MM, Davis DN. Addressing the Class Imbalance Problem in Medical Datasets. International Journal of Machine Learning and Computing. 2013:224–228.

[86] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016:770–778. Available from: http://image-net.org/challenges/LSVRC/2015/.

[87] Kerfoot E, Clough J, Oksuz I, Lee J, King AP, Schnabel JA. Left-Ventricle Quantification Using Residual U-Net. In: International Workshop on Statistical Atlases and Computational Models of the Heart; 2019. p. 371–380. Available from: https://doi.org/10.1007/978-3-030-12029-0{_}40.

[88] He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2015. p. 1026–1034.

[89] Ulyanov D, Vedaldi A, Lempitsky V. Instance Normalization: The Missing Ingredient for Fast Stylization; 2016. Available from: https://arxiv.org/abs/1701.02096.

[90] Xue Y, Xu T, Huang X. Adversarial learning with multi-scale loss for skin lesion segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging. Washington, DC., USA; 2018. p. 859–863. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp={&}arnumber=8363707{&}tag=1.

[91] Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017 jun;10541 LNCS:379–387. Available from: http://arxiv.org/abs/1706.05721.

[92] Abraham N, Khan NM. A Novel Focal Tversky loss function with improved Attention U-Net for lesion segmentation. Proceedings - International Symposium on Biomedical Imaging. 2018 oct;2019-April:683–687. Available from: http://arxiv.org/abs/1810.07842.

[93] Vinod R. Dealing with class imbalanced image datasets using the Focal Tversky Loss ; 2020. Available from: https://towardsdatascience.com/dealing-with-class-imbalanced-image-datasets-1cbd17de76b5.

[94] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings. 2016.

[95] Xia H, Sun W, Song S, Mou X. Md-Net: Multi-scale Dilated Convolution Network for CT Images Segmentation. Neural Processing Letters. 2020;51(3):2915–2927. Available from: https://doi.org/10.1007/s11063-020-10230-x.

[96] Wolterink JM, Leiner T, Viergever MA, Išgum I. Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017;10129 LNCS(2017):95–102.

[97] Cui X, Zheng K, Gao L, Yang D, Ren J. Multiscale Spatial-Spectral Convolutional Network with Image-Based Framework for Hyperspectral Imagery Classification. Remote Sensing. 2019 09;11:2220.

[98] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018 apr;40(4):834–848. Available from: http://liangchiehchen.com/projects/.

[99] Zhao D, Zhu D, Lu J, Luo Y, Zhang G. Synthetic medical images using F & BGANfor improved lung nodules classification bymulti-scale VGG16. Symmetry. 2018;10(10):1–16.

[100] Shao Q, Gong L, Ma K, Liu H, Zheng Y. Attentive CT Lesion Detection Using Deep Pyramid Inference with Multi-scale Booster. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2019;11769 LNCS:301–309.

[101] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Medical Imaging. 2015 aug;15(1):29. Available from: /pmc/articles/PMC4533825//pmc/articles/PMC4533825/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4533825/.

[102] Karimi D, Salcudean SE. Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks. IEEE Transactions on Medical Imaging. 2019 apr;39(2):499–513. Available from: http://arxiv.org/abs/1904.10030.

[103] Hsu H, Lachenbruch PA. Paired t Test . In: Encyclopedia of Biostatistics. John Wiley & Sons, Ltd; 2005. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/0470011815.b2a15112https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a15112https://onlinelibrary.wiley.com/doi/10.1002/0470011815.b2a15112.

[104] The MONAI Consortium. Project MONAI; (2020). Available from: https://monai.io/.

[105] Shen F, Zeng G. Gaussian dilated convolution for semantic image segmentation. vol. 11164 LNCS. Springer International Publishing; 2018. Available from: http://dx.doi.org/10.1007/978-3-030-00776-8{_}30.

[106] Asma-Ull H, Yun ID, Han D. Data Efficient Segmentation of Various 3D Medical Images Using Guided Generative Adversarial Networks. IEEE Access. 2020;8:102022–102031.

[107] Oner MU, Cheng YC, Lee HK, Sung WK. Training machine learning models on patient level data segregation is crucial in practical clinical applications. medRxiv. 2020. Available from: https://www.medrxiv.org/content/early/2020/04/25/2020.04.23.20076406.

[108] CoLe-CNN: Context-learning convolutional neural network with adaptive loss function for lung nodule segmentation. Computer Methods and Programs in Biomedicine. 2021 jan;198:105792.

[109] Wang S, Hu SY, Cheah E, Wang X, Wang J, Chen L, et al. U-Net using stacked dilated convolutions for medical image segmentation. arXiv. 2020.

[110] Singh VK, Romani S, Rashwan HA, Akram F, Pandey N, Sarker MMK, et al. Conditional generative adversarial and convolutional networks for X-ray breast mass segmentation and shape classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2018;11071 LNCS:833–840.

[111] Alalwan N, Abozeid A, ElHabshy AAA, Alzahrani A. Efficient 3D Deep Learning Model for Medical Image Semantic Segmentation. Alexandria Engineering Journal. 2021 feb;60(1):1231–1239.

[112] Roy R, Chakraborti T, Chowdhury AS. A deep learning-shape driven level set synergism for pulmonary nodule segmentation . Pattern Recognition Letters. 2019;123:31–38. Available from: https://doi.org/10.1016/j.patrec.2019.03.004.

[113] A cascaded dual-pathway residual network for lung nodule segmentation in CT images. Physica Medica. 2019 jul;63:112–121.

[114] Wang S, Zhou M, Liu Z, Liu Z, Gu D, Zang Y, et al. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. Medical Image Analysis. 2017 aug;40:172–183. Available from: http://dx.doi.org/10.1016/j.media.2017.06.014.

[115] Tang H, Zhang C, Xie X. NoduleNet: Decoupled False Positive Reduction for Pulmonary Nodule Detection and Segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 11769 LNCS. Springer; 2019. p. 266–274. Available from: https://doi.org/10.1007/978-3-030-32226-7{_}30.

[116] Usman M, Lee BD, Byon SS, Kim SH, il Lee B, Shin YG. Volumetric lung nodule segmentation using adaptive ROI with multi-view residual learning. Scientific Reports. 2020 dec;10(1):12839. Available from: https://doi.org/10.1038/s41598-020-69817-y.

[117] GitHub - phillipi/pix2pix: Image-to-image translation with conditional adversarial nets;. Available from: https://github.com/phillipi/pix2pix.