Semi-supervised Semantic Segmentation: Different GAN based approaches

Project (EC57004) report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Master of Technology

in

Visual Information and Embedded Systems

by

Arnab Kumar Mondal

(14EC35031)

Under the supervision of

Professor Prabir Kumar Biswas

Department of Electronics and Electrical Communication Engineering

Indian Institute of Technology Kharagpur

Spring Semester, 2018-19

May 2nd, 2019

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: May 2nd, 2019 (Arnab Kumar Mondal)

Place: Kharagpur (14EC35031)

DEPARTMENT OF ELECTRONICS AND ELECTRICAL COMMUNICATION ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Semi-supervised Semantic Segmentation: Different GAN based approaches**" submitted by **Arnab Kumar Mondal** (Roll No. 14EC35031) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Master of Technology in Visual Information and Embedded Systems is a record of bona fide work carried out by him under my supervision and guidance during Spring Semester, 2018-19.

]X[l]X[1.15c]
Professor Prabir Kumar Biswas
Date: May 2nd, 2019
Department of Electronics and Electrical Communication Engineering
Place: Kharagpur
Indian Institute of Technology Kharagpur

Kharagpur - 721302, India

Abstract

Name of the student: **Arnab Kumar Mondal** Roll No: **14EC35031**

Degree for which submitted: **Master of Technology**

Department: **Department of Electronics and Electrical Communication Engineering**

Thesis title: **Semi-supervised Semantic Segmentation: Different GAN based approaches**

Thesis supervisor: **Professor Prabir Kumar Biswas**

Month and year of thesis submission: **May 2nd, 2019**

In this work, we address the problem of segmenting medical images in scenarios where very few labeled examples are available for training. Leveraging the recent success of adversarial learning for semi-supervised segmentation, we propose multiple novel methods based on Generative Adversarial Networks (GANs) to train a segmentation model with both labeled and unlabeled images. The first proposed method prevents over-fitting by learning to discriminate between true and fake patches obtained by a generator network. We also propose a novel cycle consistency loss based approach for semi-supervised semantic segmentation. The proposed method is evaluated on the problem of segmenting retinal images of DRIVE & STARE dataset and brain MRI from the iSEG-2017 & MRBrainS 2013 datasets. Significant performance improvement is reported, compared to state-of-art segmentation networks trained in a fully-supervised manner. In addition, our work presents a comprehensive analysis of

different GAN architectures for semi-supervised segmentation, showing recent techniques yield a higher performance than conventional adversarial training approaches.

Acknowledgements

I wish to express my sincere and deepest sense of gratitude and indebtedness to my guide Dr. Prabir Kumar Biswas, Professor and Head of the Electronics & Electrical Communication Engineering Department, for his invaluable help, directions and involvement throughout this project. I would also like to thank Mr. Avisek Lahiri, a PhD student under Prof. P K Biswas for his constant support and guidance.

Contents

Declaration i

Certificate ii

Abstract iii

Acknowledgements v

Contents vi

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Related Work . . . 2
  - 1.2.1 Semi-supervised learning . . . 3
  - 1.2.2 Adversarial learning . . . 3
- 1.3 Contribution . . . 4
- 1.4 Outline . . . 5

2 Background Theory 6
- 2.1 Introduction . . . 6
- 2.2 Generative Adversarial Network . . . 7
- 2.3 Deep Convolutional GAN . . . 7
- 2.4 Conditional GANs . . . 8
- 2.5 Cycle GANs . . . 9
- 2.6 U-Net Overview . . . 10

3 Discrminator based model 12
- 3.1 Introduction . . . 12
- 3.2 Methodology . . . 12
  - 3.2.1 Center Pixel v/s Structured Prediction . . . 14
  - 3.2.2 Reduced supervision with GANs . . . 15
    - 3.2.2.1 Discriminator loss . . . 16
    - 3.2.2.2 Generator loss . . . 17

3.2.3 Complement (Bad) generator 18
3.3 Experiments with 2D medical imaging dataset 19
3.3.1 Architecture and optimisation 19
3.3.2 Datasets and Preprocessing 20
3.3.3 GAN Hacks 20
3.3.3.1 Max Pool v/s Average Pool 20
3.3.3.2 Normalization 21
3.3.3.3 Selecting layer(s) for Feature Matching 22
3.3.4 Comparison with State-of-the-art 22
3.4 Experiments with 3D medical imaging dataset 23
3.4.1 Materials 23
3.4.1.1 Dataset 25
3.4.1.2 Evaluation 26
3.4.2 Implementation Details 26
3.4.2.1 Architecture 26
3.4.2.2 Training 27
3.4.3 Detailed analysis of GAN models 29
3.4.4 Few-shot learning with FM GAN 29
3.4.5 Validation on MR Brains dataset 33

4 Generator based model **37**
4.1 Introduction 37
4.1.1 Generator based segmentation model 37
4.1.2 Methodology 37
4.2 Experiments with 2D medical imaging dataset 39
4.3 Experiments with 3D medical imaging dataset 40
4.4 Conclusion and Future Work 41

Bibliography **43**

Chapter 1

Introduction

1.1 Motivation

With the relatively new breakthrough in large scale object recognition by Krizhevsky *et al.,*(1), Convolutional Neural Networks (CNN) and 'deep learning'(DL) have achieved unprecedented success in numerous computer vision applications such as object detection(2), semantic segmentation (3), video understanding (4), visual question-answering (5) to list a few. Inspired by the flexibility of CNNs to adapt to novel computer vision problems, a recent surge of interest has been instigated among the medical image processing community to leverage the rich feature learning and representation prowess of CNNs. In recent years, CNNs have been applied in numerous medical image and video understanding pipelines such as segmenting areas of interest from medical images (6; 7; 8; 9; 10) and sequences (11), medical video understanding (12), reconstruction (13), anomaly region detection (14; 15; 16). The list is by no means exhaustive; readers are encouraged to refer to (17) for a detailed survey on applications of DL in medical image analysis.

However the success of DL comes at a price. CNN models are significantly complex with millions of trainable parameters. For example, popular architectures such as AlexNet (1) and VGG-Net(18) have 60 million and 138 million parameters respectively. Such gigantic deep architectures easily overfit on small training datasets with low training error but manifests high test error. Curating manually annotated

dataset is both time consuming and costly. Even though for natural computer vision problems the current trend is to annotate large scale data with mechanical turks (19), annotating medical data often requires domain specific experts. This instigates the need for methods to train CNNs with limited amount of annotated data. Recent regularization techniques such as dropout (20) and batch normalization (21) have shown promise in preventing over fitting; however a small dataset with regularization during training can easily lead to under fitting, wherein, during the training phase itself, a CNN is unable to approximate the input to output functional mapping appreciably, thereby manifesting high error rates on both training and testing data. Fine-tuning a pre-trained CNN (in most cases pre-trained for object recognition on ImageNet) for specific medical imaging tasks (22) is the current trend to train a CNN with limited annotated data. Though promising, fine-tuning methods train a CNN by only annotating a fraction of available data, while the remaining unannotated data remains unused.

Semi-supervised learning approaches alleviate the need for large sets of labeled samples by exploiting available non-annotated data. In such approaches, only a limited number of samples with strong annotations are provided. A good generalization can however be achieved by considering unlabeled samples, or samples with weak annotations like image-level tags (23; 24; 25; 26), bounding boxes (27; 28) or scribbles (29; 25), during training. Recently, approaches based on adversarial training, and in particular Generative Adversarial Networks (GANs) (30), have shown great potential for improving semantic segmentation in a semi-supervised setting (31; 32).

1.2 Related Work

Our method draws on recent successes of deep learning methods for semantic segmentation, in particular semi-supervised and few-shot learning approaches based on adversarial training.

1.2.1 Semi-supervised learning

Several semi-supervised deep learning methods have been proposed for image segmentation (33; 34; 35; 36). A common strategy, based on the principle of self-training, involves updating network parameters and segmentation alternatively until convergence (33). However, if initial class priors given by the network are inaccurate, segmentation errors can occur and be propagated back to the network which then re-amplifies these errors. Various techniques can be used to alleviate this problem, including model-based (37) or data-based (38; 36) distillation, which aggregate the prediction of multiple teacher models or a single teacher trained with multiple transformed versions of the data to learn a student model, and employing attention modules (35). Yet, these approaches are relatively complex, as they require to train multiple networks, and are thus not suitable when very few training samples are available. Another popular approach consists in embedding the network's output or internal representation in a manifold space, such that images having similar characteristics are near to each other (34). An important limitation of this approach is its requirement for an explicit matching function, which may be hard to define in practice.

1.2.2 Adversarial learning

Adversarial learning has also shown great promise for training deep segmentation models with few strongly-annotated images (31; 32; 39; 40). An interesting approach to include unlabeled images during training is to add an adversarial network in the model, which must determine whether the output of the segmentation network corresponds to a labeled or unlabeled image (39; 40). This encourages the segmentation network to have a similar distribution of outputs for images with and without annotations, thereby helping generalization. A potential issue with this approach is that the adversarial network can have a reverse effect, where the output for annotated images becomes growingly similar to the incorrect segmentations obtained for unlabeled images. A related strategy uses the discriminator to predict a confidence map for the segmentation, enforcing this output to be maximum for annotated images (32). For unlabeled images, areas of high confidence are used to update the segmentation network in a self-teaching manner. The main limitation of this approach

is that a confidence threshold must be provided, the value of which can affect the performance. Up to date, only a single work has applied Generative Adversarial Networks (GANs) for semi-supervised segmentation (31). However, it focused on 2D natural images, whereas the current work targets 3D multi-modal medical volumes. Generating and segmenting 3D volumes brings additional challenges, such as computational complexity and over-fitting.

1.3 Contribution

Our work addresses the problem of segmenting images from a semi-supervised learning perspective. We leverage the recent success of GANs to train a deep model with a highly-limited training set of labeled images, without sacrificing the performance of full supervision. The extra information is learned from the unlabeled dataset. We chose challenging medical imaging datasets because it's both expensive and time consuming process to obtain annotated samples. The main contributions of this work can be summarized as:

1. We provide a novel GAN based approach for semi-supervised semantic segmentation and later improve the model by incorporating the concept of badGAN. We tested the basic approach with 2D medical images of retina from DRIVE and STARE dataset and then moved to more challenging 3D multimodal medical images from brain MRI segmentation challenges like iSEG 2017 and MR Brains 2013.

2. We provide another novel approach based on Cycle Consistency based Generative Adversarial Network which further improves the performance. This method uses the segmenter model as generator unlike the previous one which uses it as a discriminator.

3. We demonstrate that the proposed approach to significantly outperform state-of-art segmentation networks like UNet when very few labeled training samples are available using semi-supervised learning, and to achieve an accuracy close to that of full supervision.

4. A comprehensive analysis of different GAN architectures for semi-supervised segmentation, where we show more recent techniques have a higher performance than conventional adversarial training approaches.

1.4 Outline

This thesis is mainly divided into 4 chapters. It starts with a brief introduction which includes motivation for this work, a brief literature review on related work and finally summarizes our contribution. The second chapter starts with basic basic background theories necessary to understand this work properly. The third chapter describes the first part of this master's thesis where a Discriminator based model is proposed. It starts with the methodology and then provides detailed discussion on the architecture used and the experiments we did to eexplore the working and substantiate the utility of our model. We run experiments on both 2D and 3D medical imaging datasets to establish this technique for all kinds of datasets. Chapter 4 introduces a new generator based model which showed to beat the discriminator based models and a detailed comparison between all the models is provided. Finally we conclude this entire work in that chapter.

Chapter 2

Background Theory

2.1 Introduction

The basic paradigm of deep learning is to discover rich, hierarchical models that represent probability distributions over various kinds of data such as natural images, audio waveforms containing speech, and symbols in natural language corpora. These models can be of two kinds - discriminative or generative. Discriminative models, which have been successfully used in many deep learning problems, usually map a high-dimensional input to a class label. They essentially find the conditional probability of the class label given the input. Deep generative models, on the other hand, aim the learn the joint probability of the input and the class labels, that is, the underlying probability distribution of the data generation process.

In the adversarial nets framework, the generative model is pitted against an adversary: a discriminative model that learns to determine whether a sample is from the model distribution or the data distribution. Competition in this game drives both teams to improve their methods until the generated samples are indistinguishable from the original data.

2.2 Generative Adversarial Network

Generative adversarial network (GAN) (30) presents a two player min-max game between a generator (G) and discriminator (D) network. The idea is to simultaneously train the D and G networks. G is trained to map random vectors $z \in R^Z$ to synthetic image vector, $\tilde{x} = G(z)$. The objective of D network is to distinguish between real examples, $x \sim p_{data(x)}$, from synthetic examples, $G(z) \sim p_{G(z)}$ generated by G. D(x) represents the probability that a sample x belongs to original data distribution. Gradient of output of D with respect to its input is used by G to update its own parameters. Specifically, D and G play a two player min-max game with value function V(G,D):

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

Though, initially GAN was proposed as an approximate sampler from original data distribution, the concept of adversarial learning have been successfully applied over diversified computer vision applications such as image super resolution, image impainting, image-to-image translation and video frame prediction to name a few.

2.3 Deep Convolutional GAN

In recent years, supervised learning with convolutional networks (CNNs) has seen huge adoption in computer vision applications. Comparatively, unsupervised learning with CNNs has received less attention. We can use convolutional networks in the GAN framework to give rise to a new class of networks called deep convolutional generative adversarial networks (DCGANs)(41). The basic framework of a generative adversarial network remains the same, however the generator and discriminator networks are now convolutional networks.

The generator network is a deconvolutional neural network, composed of a number of deconvolutional layers. Each such layer in the hierarchy groups information from the layer beneath to form more complex features that exist over a larger scale in

the image. The latent noise vector $\boldsymbol{z}$ goes through a series of such deconvolution operations to give an image.

The discriminator network is a convolutional neural network, composed of a series of convolutional layers. Each such layer in the hierarchy applies learned spatial filters to the layer beneath to systematically form more abstract features. The discriminator network takes an image as input and after passing it through a series of convolution operations outputs a vector of probabilities corresponding to each output class.

2.4 Conditional GANs

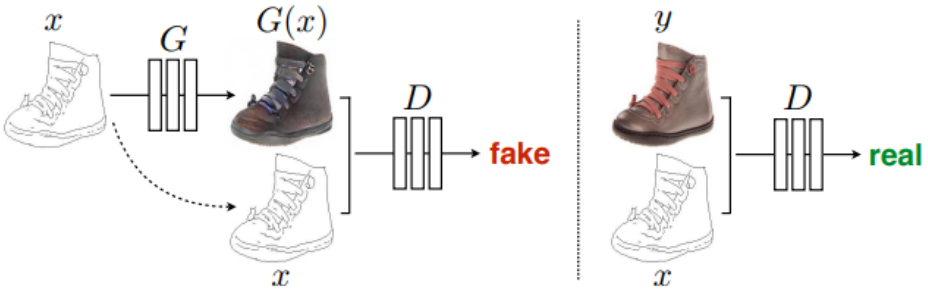


FIGURE 2.1: Training a conditional GAN to map edges → photo. The discriminator, D, learns to fool the discriminator. Unlike an unconditional GAN, both the generator and discriminator observe the input edge map.

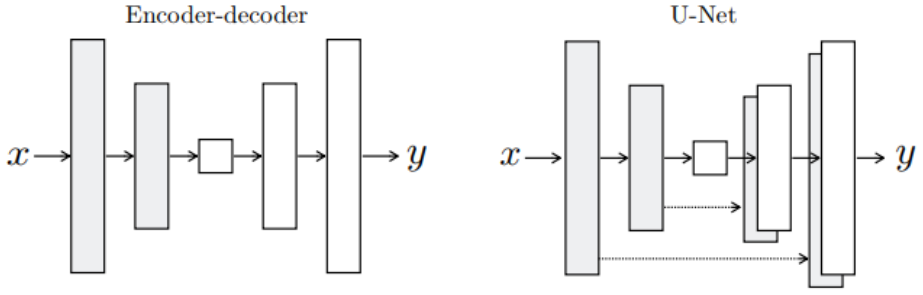


FIGURE 2.2: Two choices for the architecture of the generator. The "U-Net" (42) is an encoder-decoder with skip connections between mirrored layers in the encoder and decoder stacks.

GANs are generative models that learn a mapping from random noise vector z to output image y, $G : z \rightarrow y$ (30). In contrast, conditional GANs (43) learn a mapping from observed image x and random noise vector z, to y, $G : x, z \rightarrow y$. The generator G is trained to produce outputs that cannot be distinguished from "real" images by an adversarially trained discriminator, D, which is trained to do as well as possible at detecting the generator's "fakes". This training procedure for image translation with conditional GANs is shown in Figure 2.1.Two of the generator networks which are generally used for this task are shown in Figure 2.2.

2.5 Cycle GANs

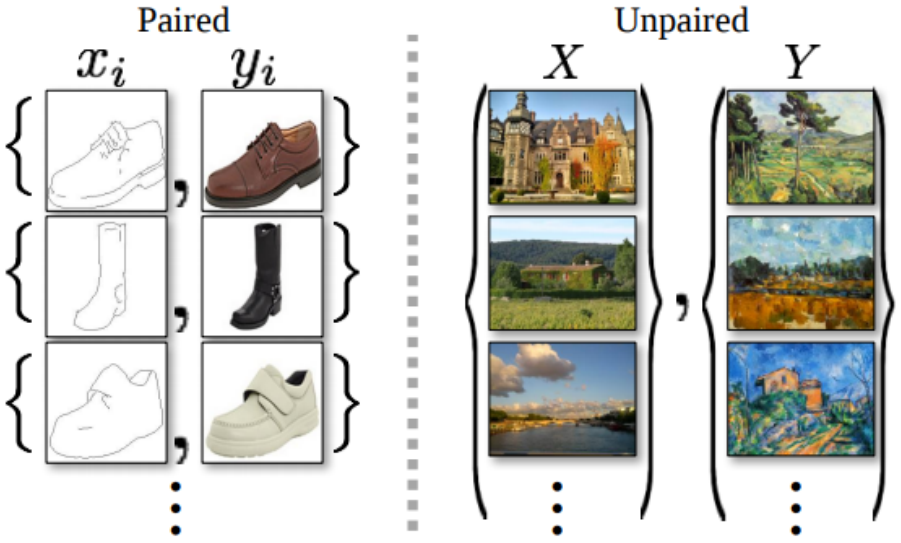


FIGURE 2.3: Paired training data (left) consists of training examples $\{x_i, y_i\}_{i=1}^{N}$, where the correspondence between x_i and y_i exists [22]. We instead consider unpaired training data (right), consisting of a source set $\{x_i\}_{i=1}^{N}(x_i \epsilon X)$ and a target set $\{y_j\}_{j=1}^{N}(y_i \epsilon Y)$, with no information provided as to which x_i matches which y_j .

In CycleGANs(44), the goal is to learn mapping functions between two domains X and Y given training samples $\{x_i\}_{i=1}^{N}$ where $x_i \epsilon X$ and $\{y_j\}_{j=1}^{M}$ where $y_j \epsilon Y$. The training images to obtain the mapping need not be paired which makes it suitable for applications which only have unpaired images as shown in Figure 2.3. We denote the data distribution as $x \sim p_{data(x)}$ and $y \sim p_{data(y)}$. As illustrated in Figure

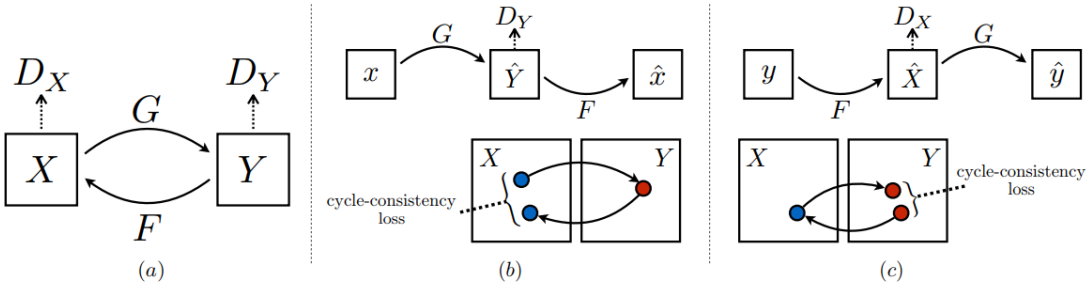


FIGURE 2.4: (a) The model contains two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators D_Y and D_X. D_Y encourages D to translate X into outputs indistinguishable from domain Y , and vice versa for D_X and F. To further regularize the mappings, two cycle consistency losses are introduced that capture the intuition that if we translate from one domain to the other and back again we should arrive at where we started: (b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$

2.4, the model includes two mappings $G : X \rightarrow Y$ and $F : Y \rightarrow X$. In addition, two adversarial discriminators are included D_X and D_Y , where D_X aims to distinguish between images x and translated images $F(y)$; in the same way, DY aims to discriminate between y and $G(x)$. The objective contains two types of terms: adversarial losses (30), which is similar to the used in Conditional GANs, for matching the distribution of generated images to the data distribution in the target domain; and cycle consistency losses (45) to prevent the learned mappings G and F from contradicting each other.

2.6 U-Net Overview

The main idea is to supplement a usual contracting network by successive layers, where pooling operators are replaced by upsampling operators. Hence, these layers increase the resolution of the output. In order to localize, high resolution features from the contracting path are combined with the upsampledoutput. A successive convolution layer can then learn to assemble a more precise output based on this information.

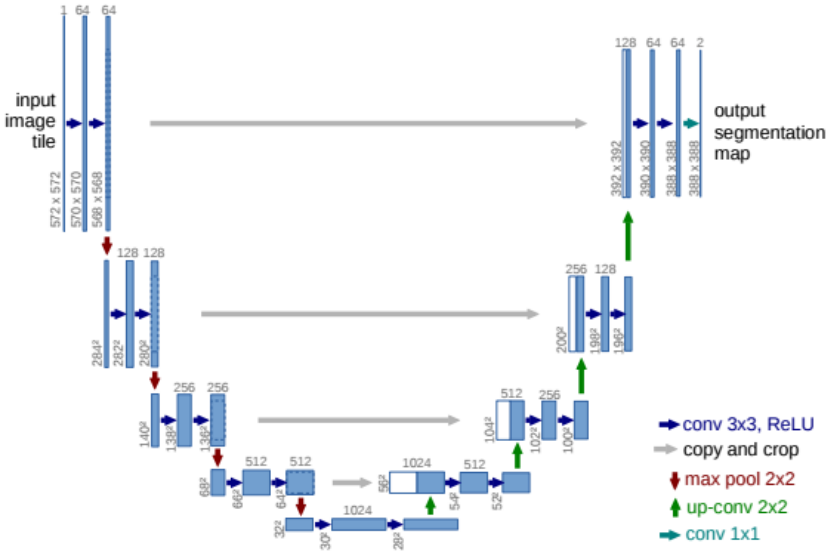


FIGURE 2.5: General U-Net Architecture

One important thing in this architecture is that in the upsampling part they have also a large number of feature channels, which allow the network to propagate context information to higher resolution layers. As a consequence, the expansive path is more or less symmetric to the contracting path, and yields a u-shaped architecture. The network does not have any fully connected layers and only uses the valid part of each convolution, i.e., the segmentation map only contains the pixels, for which the full context is available in the input image.

For training this model the sum of cross entropy loss of every pixel of output segmented map is minimized. Once the network is trained it has been shown to give best in class results in most medical image segmentation problems. Later we will compare its performance with our model.

Chapter 3

Discrminator based model

3.1 Introduction

In this chapter we provide detailed methodology used by the discriminator based for semi-supervised semantic segmentation. We test our model in 2D medical imaging datasets of retina and 3D medical imaging datasets of Brain MRI images and provide detailed analysis of our results. We show all the theory with respect to the 3D model which can be easily adapted to the 2D one. We extend our GAN based model to badGANs and show how it performs differently with 2D and 3D datasets.

3.2 Methodology

The proposed architecture for the semi-supervised segmentation of 3D medical images is illustrated in Figure 3.1. In a standard segmentation model such as U-Net (42), fully-annotated images are typically employed to train the network using a pixel-wise loss function like cross-entropy. As mentioned above, this is not possible in our case since the number of annotated images for training is highly limited. As in other semi-supervised segmentation approaches, we alleviate this problem by also incorporating unlabeled images in the training process. However, unlike these methods, we also make use of synthetic (i.e., *fake*) images generated by a GAN.

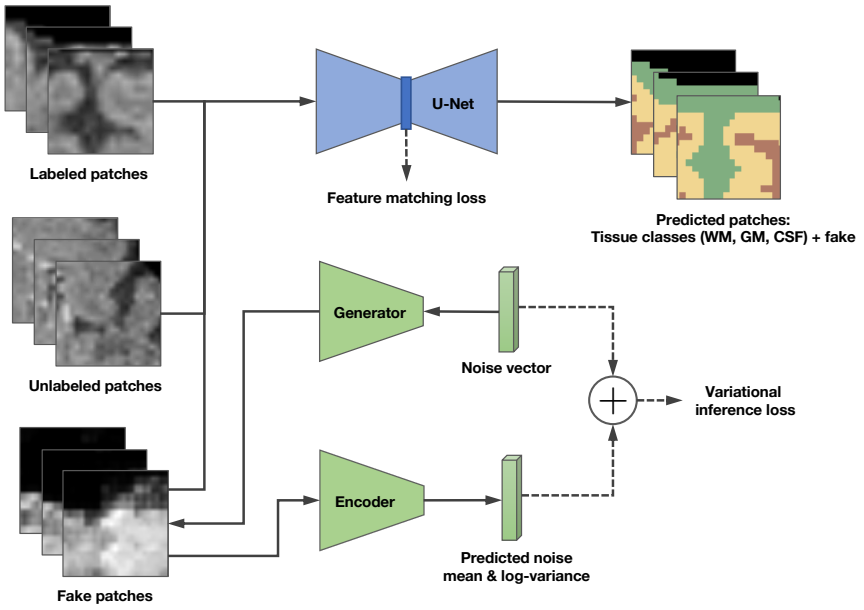


FIGURE 3.1: Schematic explaining the working of our model. The model contains three networks which are trained simultaneously.

To include labeled, unlabeled and fake images during training, we extend the classification approach of Dai et al. (46) to segmentation. In this GAN-based approach, a generator network is used to produce realistic fake examples and a discriminator to distinguish these fake examples from true data. Instead of predicting K classes, as in standard methods, the model predicts $(K+1)$ classes, where the additional class corresponds to fake examples. However, as we show in following subsections, this formulation can be recast back into a K-class problem using a simple re-parametrization trick. This overall strategy helps the model give plausible predictions for unlabeled true data by restricting its output for fake examples.

For adapting this model to the segmentation of 3D multi-modal images, several changes must be made. During training, 3D images must be processed in smaller sub-regions (i.e., *patches*) to deal with the much greater memory and computational requirements compared to 2D images. While training patches have a fixed size, test images may have arbitrary size. To address this issue, we make the segmentation network to be fully-convolutional (47). Another challenge comes from the generation of fake patches. Standard techniques for training GANs may lead to instability and poor results, especially in the case of semi-supervised learning (48). This problem is even more significant in the case of 3D multi-modal patches, whose distribution is

harder to estimate with a parametric model. In addition, although generated patches must be realistic-looking, they should be sufficiently different from true unlabeled patches, otherwise the wrong information will be learned (46).

The following subsections provide a more detailed description of the proposed method. We start by giving a general formulation of generative adversarial networks (GANs). Then, we show how GANs can be used to include unlabeled and fake images in a semi-supervised segmentation setting. Finally, we explain how the standard GAN model is modified to fit our problem setting.

3.2.1 Center Pixel v/s Structured Prediction

There are two major paradigms for patchwise segmentation of medical images, namely center-pixel prediction (CP) and structured prediction (SP).

Let P be the domain of sampled patches from the dataset such that any $p \sim P \in \mathbb{R}^{H\times W\times D\times 1}$, where $H \times W \times D$ is the resolution of the patches. Also, let Y be the corresponding label space for P, such that for a given patch, x_p, we have its corresponding label, $y_p \in \mathbb{R}^{H\times W\times D}$. $y_p^{i,j,k}$ is the label information at location (i, j, k) for patch x_p. In case of center pixel prediction, the objective is to learn a parametrized (θ_C) functional mapping, $f_{\theta_C} : P \Rightarrow \mathbb{R}^{1\times 1}$. Essentially this means that given an image patch, the function returns a single scalar value to predict the probability of the center pixel of that patch to belong to foreground or 'vessel' class. θ_C is optimized according to,

$$\theta_C^* = \underset{\theta_C}{arg\,min} \; -\sum_{p=1}^{m} y_p^{(H/2,W/2,D/2)} \log(f_{\theta_C}(x_p)) + \; (1 - y_p^{(H/2,W/2,D/2)}) \log(1 - f_{\theta_C}(x_p)) \tag{3.1}$$

This was the procedure we followed in (49).

In contrast, structured prediction learns a parametrized function, $f_{\theta_S} : P \Rightarrow \mathbb{R}^{H\times W\times D}$. θ_S is optimized as,

$$\theta_S^* = \underset{\theta_C}{arg\,min} \; -\sum_{p=1}^{m}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{D} y_p^{(i,j,k)} \log(f_{\theta_C}(x_p))$$

$$+ \ (1 - y_p^{(i,j,k)}) \log(1 - f_{\theta_C}(x_p)) \tag{3.2}$$

In this methodology, for a given image patch, the function simulteneously predicts the probability of all the pixels in the patch belonging to the 'vessel' class, instead of just the center pixel.

3.2.2 Reduced supervision with GANs

Consider a standard CNN-based model for segmenting a 3D image $\mathbf{x}_{H\times W\times D}$ into regions defined by $\mathbf{y}_{H\times W\times D}$. This model takes $\mathbf{x}$ as input and outputs a K-dimensional vector of logits $[l_{i,1}, \ldots, l_{i,K}]$, where K is the number of classes labels and i is the index of image voxels. This output can be turned into class probabilities by applying the softmax function:

$$p_{\text{model}}(y_i = j \,|\, \mathbf{x}) \ = \ \frac{\exp(l_{i,j})}{\sum_{k=1}^{K} \exp(l_{i,k})}. \tag{3.3}$$

In a fully-supervised setting, the model is typically trained by minimizing a segmentation loss function, for instance, the cross-entropy between the true labels and the model's predicted probabilities.

As shown in Fig. 3.1, the proposed model extends the standard full-supervision approach by incorporating unlabeled data and samples from the generator G during training. Toward this goal, we label generated samples with a new class $y_i = K{+}1$ and thus increase the dimension of the segmentation model's output to $H{\times}W{\times}D{\times}(K{+}1)$. In this new formulation, $p_{\text{model}}(y_i = K+1 \,|\, \mathbf{x})$ is the probability that voxel i of input $\mathbf{x}$ is fake. Moreover, to learn the basic structure of images from unlabeled data, we constraint the output to correspond to one of the K classes of real data, which can be done by maximizing

$$\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})} \sum_{i=1}^{H\times W\times D} \log \, p_{\text{model}}(y_i \in \{1, \ldots, K\} \,|\, \mathbf{x}). \tag{3.4}$$

With this, we can now define the loss functions used for training the discriminator and generator networks.

3.2.2.1 Discriminator loss

Suppose have a similar number of labeled, unlabeled and fake images, so that each type of images has equal importance in training. Our discriminator loss function can be defined as the sum of three terms:

$$L_{\text{discriminator}} = L_{\text{labeled}} + L_{\text{unlabeled}} + L_{\text{fake}}. \tag{3.5}$$

The loss for labeled images is the same as in standard segmentation networks. In this work, we consider the mean cross-entropy:

$$L_{\text{labeled}} = -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{\text{data}}(\mathbf{x},\mathbf{y})} \sum_{i=1}^{H\times W\times D} \log\, p_{\text{model}}(y_i \,|\, \mathbf{x},\, y_i < K+1). \tag{3.6}$$

In the case of unlabeled images, we maximize the term in Eq. (3.4), which is the same as minimizing

$$L_{\text{unlabeled}} = -\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})} \sum_{i=1}^{H\times W\times D} \log\, \Big[1 - p_{\text{model}}(y_i = K+1 \,|\, \mathbf{x})\Big] \tag{3.7}$$

Finally, for generated images, we impose each pixel of an input patch to be predicted as fake, and define the loss as

$$L_{\text{fake}} = -\mathbb{E}_{z\sim \text{noise}} \sum_{i=1}^{H\times W\times D} \log\, p_{\text{model}}\Big(y_i = K+1 \,|\, G_{\theta_G}(z)\Big) \tag{3.8}$$

In (48), it was shown that the optimal strategy for minimizing Eq. (3.5) is to have $\exp[l_{i,j}] = c_i(\mathbf{x}) \cdot p(y_i = j,\, \mathbf{x}), \quad \forall j < K+1$ and $\exp[l_{i,K+1}] = c_i(\mathbf{x}) \cdot p_G(\mathbf{x})$, where $c_i(\mathbf{x})$ is an undetermined scaling function for the i-th pixel. It was also found that the having $K+1$ outputs is an over-parameterized formulation, since subtracting a general function $f(\mathbf{x})$ from each logit does not change the output of the softmax. By using the logit of the fake class $l_{i,K+1}$ as subtracted function, we obtain $\Big[(l_{i,1} - l_{i,K+1}), \ldots, (l_{i,K} - l_{i,K+1}), 0\Big]$, and thus have only K effective (i.e., non-zero) outputs. Employing these "normalized" logits in the softmax of Eq. (3.3) then leads to the

following modified loss functions:

$$L_{\text{labeled}} = -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{\text{data}}(\mathbf{x},\mathbf{y})} \sum_{i=1}^{H\times W\times D} \log\, p_{\text{model}}(y_i \,|\, \mathbf{x}) \tag{3.9}$$

$$L_{\text{unlabeled}} = -\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})} \sum_{i=1}^{H\times W\times D} \log\left[\frac{Z_i(\mathbf{x})}{Z_i(\mathbf{x})+1}\right] \tag{3.10}$$

$$L_{\text{fake}} = -\mathbb{E}_{\mathbf{z}\sim \text{noise}} \sum_{i=1}^{H\times W\times D} \log\left[\frac{1}{Z_i(G_{\theta_G}(\mathbf{z}))+1}\right] \tag{3.11}$$

where $Z_i(x) = \sum_{k=1}^{K} \exp[l_{i,k}(\mathbf{x})]$.

In summary, the idea is to plug a standard state-of-the-art segmentation model in the discriminator of the proposed network, where the labeled component of the loss L_{labeled} remains unchanged (i.e, cross-entropy), and introduce two extra terms, the unlabeled term $L_{\text{unlabeled}}$ and the fake term L_{fake}, which are analogous to the two components of a discriminator loss in standard GANs.

3.2.2.2 Generator loss

The most common strategy for training the generator consists in maximizing the L_{fake} loss of Eq. (3.8). However, as demonstrated in (48), this can lead to instability and poor performance in the case of semi-supervised learning. Following these results, we instead adopt the Feature Matching (FM) loss for the generator, which is more suited to our problem. In FM, the goal of the generator is to match the expected value of features $f(\mathbf{x})$ in an intermediate layer of the discriminator:

$$L_{\text{generator}} = \left\| E_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}\, f(\mathbf{x}) \;-\; E_{z\sim \text{noise}}\, f(G_{\theta_G}(\mathbf{z})) \right\|_2^2 \tag{3.12}$$

In this work, $f(\mathbf{x})$ contains the activations of the second last layer of the encoding path in our model. In preliminary experiments, we found this choice to give slightly higher performance than using the encoder's last layer.

3.2.3 Complement (Bad) generator

In semi-supervised learning, having a good generator can actually deteriorate performances since in this case unlabeled and fake images cannot be separated. It is therefore desirable to have a generator that can generate samples outside the true data manifold, which is called a *complement* (or *bad*) generator (46).

The FM generator loss, described in the previous section, works better than standard training approaches in a semi-supervised setting because it performs distribution matching in a weak manner. However, it may still face two significant problems.

First, since an FM-based generator can assign a significant amount of probability mass inside the support, an optimal discriminator will incorrectly predict samples in that region as fake. Secondly, as FM only matches first-order statistics, the generator might end up with a trivial solution, for example, it can collapse to mean of unlabeled features. The collapsed generator will then fail to cover some areas between manifolds. Since the discriminator is only well-defined on the union of the data supports of p and p_G, the prediction result in such gaps is under-determined.

The first problem is less likely in our case, since multi-modal 3D patches are complex structures to generate and, thus, it is more probable for the FM generator to sample images outside the true data manifold. To deal with the second problem, we can increase the entropy of the generated distribution by minimizing a modified loss for the generator:

$$L_{generator} \;=\; -H(p_G) \;+\; \Big\| E_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}\, f(\mathbf{x}) \;-\; E_{\mathbf{x}\sim p_G}\, f(\mathbf{x}) \Big\|_2^2 \tag{3.13}$$

As mentioned in (46), this complex loss function can be optimized using a variational upper bound on the negative entropy (50):

$$-\,H\Big(p_G(\mathbf{x})\Big) \leq -\,\mathbb{E}_{\mathbf{x},\mathbf{z}\sim p_G} \log q(\mathbf{z}\,|\,\mathbf{x}). \tag{3.14}$$

In this formulation, q is defined as a diagonal Gaussian with bounded variance, i.e. $q(\mathbf{z}\,|\,\mathbf{x}) \;=\; \mathcal{N}(\mu(\mathbf{x}),\, \sigma^2(\mathbf{x}))$, with $0 \leq \sigma(\mathbf{x}) \leq \theta$, where $\mu(\cdot)$ and $\sigma(\cdot)$ are neural networks.

The overall architecture of the complement generator is illustrated in Fig. 3.1. As presented above, the FM loss uses features from the second layer of the discriminator (i.e., the U-Net segmentation network). Moreover, the fake image generator is paired with an encoder which learns a reverse mapping from generated images to corresponding noise vectors. All components the architecture are trained simultaneously in an end-to-end manner.

3.3 Experiments with 2D medical imaging dataset

3.3.1 Architecture and optimisation

The U-Net model is used as a discriminator which consists of an encoder section which creates a bottleneck starting from original image patch with a series of convolutional layers with dropout and pooling. In the decoder section, we gain back original resolution by upsampling and deconvolutional layers. In between, there are skip connections to concatenate lower and higher order features and easier flow of gradients. Unless otherwise stated, we use dropout(20) with keep probability of 0.8. Leaky Relu activation is used after every convolution with negative gradient of 0.2. For realizing the generator, we follow the principles in (51). First, the 100D z vector is passed through a linear layer and reshaped to a spatial resolution of $W/8 \times W/8$, where, $W \times W$ is the input patch resolution to the discriminator(W = 48 in our case). Then, we follow up with three transposed convolutional layers(also commonly known as deconvolutional layers) (52) to increase resolution by 2X in each step to finally reach $W \times W$. Each layer is followed by Relu non linearity except the last layer which used tanh non linearity to scale output values in the range [-1, 1].

We use mini batch stochastic gradient descent optimization with Adam optimizer(53) to train both generator and discriminator network. Learning rate for both the network are set to 10^{-4}. Batch size is kept at 64 and training usually progresses for 50 epochs in about 10 hours.

3.3.2 Datasets and Preprocessing

We conduct experiments on DRIVE (54) and STARE[1] datasets. DRIVE dataset has a clear demarcation of training and test set with 20 images in each category. Such breakup is not provided on STARE. Following recent practice (55), we follow a 1-held-out strategy, where we randomly select 1 image for testing and remaining 19 as train set. Results reported on STARE are average of 20 such trials.

The retinal images were converted to gray scale. It has been shown in (56) that the green channel in color fundus imaging is most discriminative in segmenting blood vessels. Following this, the green channel is given more weight in RGB to gray scale conversion. The contrast of the fundus images are improved using Contrast Limited Adaptive Histogram Equalization (CLAHE) and effect of non-uniform illumination is mitigated. Further, Gamma adjustment improves segmentation performance. Patches of resolution, 48×48 are then extracted from the images.

3.3.3 GAN Hacks

Finding Nash Equlibrium in a zero-sum minmax game such as in GANs is difficult (often resulting in oscillations) with stochastic gradient descent updates. This is a burning issue within GAN community. In this section, we present a detailed ablation study on various aspects of stabilizing GAN training with a basic U-Net as a baseline. Since The U-Net model is an essential component of many recent medical imaging applications, our findings in this section can serve as a guideline for any GAN based application which deploys U-Net at its core. In Table **??** we report the AUC on DRIVE test set by training with 1K labeled samples with Feature Matching and vanilla GAN and also comparing the efficacy of different GAN stabilization techniques. Similar trends were also observed for STARE.

3.3.3.1 Max Pool v/s Average Pool

In an encoder-decoder architecture like the U-Net, it is common to use Max Pool operations for spatial reduction of intermediate feature layers. This results in sparse

[1]Available at: http://cecas.clemson.edu/ ahoover/stare/

gradient operations which have been shown to hamper GAN training(51). Specifically, a Max Pool operator, $M^{W\times W}(\cdot)$, operating on a receptive field of $W \times W$ resolution of a given feature map location, $F(x, y)$, results in finding the max value in $W \times W$ neighborhood.

$$M^{W\times W}(F(x,y)) = max\{F(x-i, y-j)| \quad i,j \in \{-W/2, .., W/2\}\} \tag{3.15}$$

Instead of using Max Pool, we benefited by using Average Pooling, which also achieves spatial reduction but with dense gradient operations. In line in notations with Eq. 3.15, we define Average Pool operator, $A^{W\times W}(\cdot)$ as,

$$A^{W\times W}(F(x,y)) = \frac{1}{W^2}\sum_{i=-W/2}^{W/2}\sum_{j=-W/2}^{W/2} F(x-i, y-j) \tag{3.16}$$

3.3.3.2 Normalization

Normalization of intermediate activations/weights play a decisive role in success of training GANs. With the onset of 'DCGAN' (51), BatchNormalization (BN) (57) has become the de facto choice of normalizing weights of a deep network for GAN training. While BN indeed speeds of training of GANs, recent works, specially in the domain of style transfer, recommends the use of Instance Normalization (IN) (58) for better training of GANs. Our initial experiments also manifested better efficacy of IN over BN. IN + Max Pool did not show any significant improvement over only Max Pool which bolsters the fact that sparse gradient operations such as Max Pool are detrimental for GAN training. We further improved the performance by adopting the recent Weight Normalization (WN) technique proposed by Salimans *et al.* (59)[2]. For a linear layer,

$$\mathbf{y} = \mathbf{W^T X} + \mathbf{b}, \tag{3.17}$$

[2]Implementation available at: https://github.com/TimSalimans/weight_norm

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{W} \in \mathbb{R}^{n \times m}$, WN re-parametrizes $\mathbf{W}$ with $\mathbf{V} \in \mathbb{R}^{n \times m}$ and a trainable scalar, $g \in \mathbb{R}^m$ according to,

$$\mathbf{w_i} = \frac{\mathbf{g_i}}{||\mathbf{v_i}||_2} \cdot \mathbf{v_i}. \tag{3.18}$$

As shown in (59), decoupling of norm of the weight vector, g, from the direction of the weight vector, $\frac{\mathbf{v}}{||v||}$, helps in faster(and better) convergence of stochastic gradient descent optimization. Our GAN training also benefited using WN.

3.3.3.3 Selecting layer(s) for Feature Matching

In their original implementation, [3] Salimans *et al.* (60) used the penultimate layer of the discriminator for matching features from a batch of real and fake samples. We hypothesize that for low level vision tasks, matching features from such deeper layers of a network is not a prudent approach. For cases in which the end task is simple classification, such as in (60), it makes sense to only focus on higher order features from deeper parts of the network. Features essential for classification are agnostic to local perturbations. But in our case, the fully convolutional discriminator is responsible for semantic segmentation - assigning class label to each pixel of a patch. This requires low level information along with high level features. In fact, our initial experiments with feature matching on the penultimate layer of discriminator yielded the worst AUC performance. It appears that selecting the extremely shallow or deep layer hurts the performance. It is prudent to match intermediate layers for our low level vision task. The proof of concepts learnt so far on 2D medical images were also extended on 3D medical images experiments, unless otherwise stated.

3.3.4 Comparison with State-of-the-art

In Tables 4.1 and 4.2 we compare performance of our U-Net GAN model with the vanilla U-Net model [4]. At full supervision with 60K labeled samples, the vanilla U-Net achives AUC of 0.97 on DRIVE and 0.96 on STARE and thus U-Net serves

[3] Available at: https://github.com/openai/improved-gan

[4] Implementation adapted from https://github.com/orobix/retina-unet

TABLE 3.1: Comparison of competing supervised and semi supervised methods on DRIVE dataset.

Genre	Method	Annotated Patches			
		0.5K	1K	3K	10K
Supervised	Dasgupta *et al.*(61)	0.85	0.87	0.89	0.92
	Liskowski *et al.*(55)	0.83	0.84	0.87	0.92
	U-Net	0.89	0.90	0.92	0.95
Semi Supervised	Lahiri *et al.* (49)	0.82	0.84	0.85	0.93
	Proposed (SP)	0.92	0.94	0.96	0.97

as a very strong baseline for supervised training. At very low number of annotated patches, our model consistently outperforms U-Net across both datasets. Also, we gain distinct gain over our previous semi-supervised framework (49). We also compared against two contemporary benchmark supervised benchmark models of (61) and (55) and achieved consistent gains at different levels of annotation on both datasets. The current work thus sets up a new benchmark for such low annotation retinal vessel segmentation across two real life fundus datasets.

TABLE 3.2: Comparison of competing supervised and semi supervised methods on STARE dataset.

Genre	Method	Annotated Patches			
		0.5K	1K	3K	10K
Supervised	Dasgupta *et al.*(61)	0.82	0.84	0.87	0.91
	Liskowski *et al.*(55)	0.84	0.86	0.89	0.93
	U-Net	0.86	0.89	0.90	0.94
Semi Supervised	Lahiri *et al.* (49)	0.80	0.81	0.83	0.90
	Proposed (SP)	0.90	0.92	0.94	0.96

3.4 Experiments with 3D medical imaging dataset

3.4.1 Materials

The proposed model is evaluated on the challenging tasks of segmenting infant and adult brain tissue from multi-modal 3D magnetic resonance images (MRI). The goal of our experiments is two-fold. First, we assess our GAN-based model in a few-shot learning scenario, where only a few training subjects are provided. Our objective is to provide performance similar to that of full-supervision, while using only 1 or 2

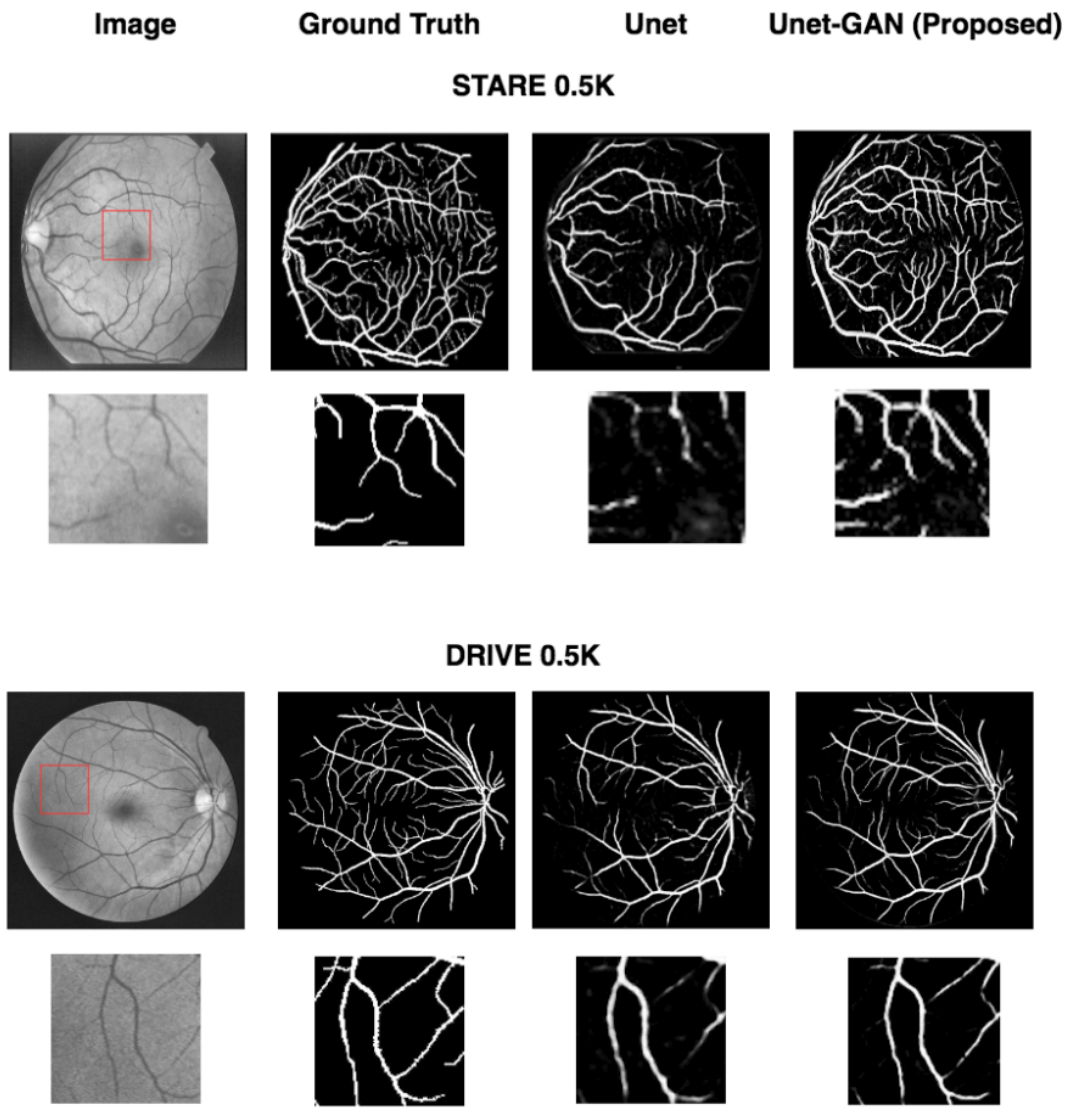


FIGURE 3.2: Some sample visualizations of segmented vessels on DRIVE and STARE dataset at 0.5K patch annotation budget. For each figure, we also show a zoomed up section. The effect is more pronounced on STARE dataset which consists of data from patient group with various opthalmic disorders.

training subjects. Second, since the application of GANs to semi-supervised learning, and particularly to segmentation, is a new topic, we conduct experiments to measure to impact of various GAN techniques (e.g., feature matching, complementary generator, etc.) on segmentation accuracy. Before presenting results, we give details on the dataset, evaluation metrics and implementation used in the experiments.

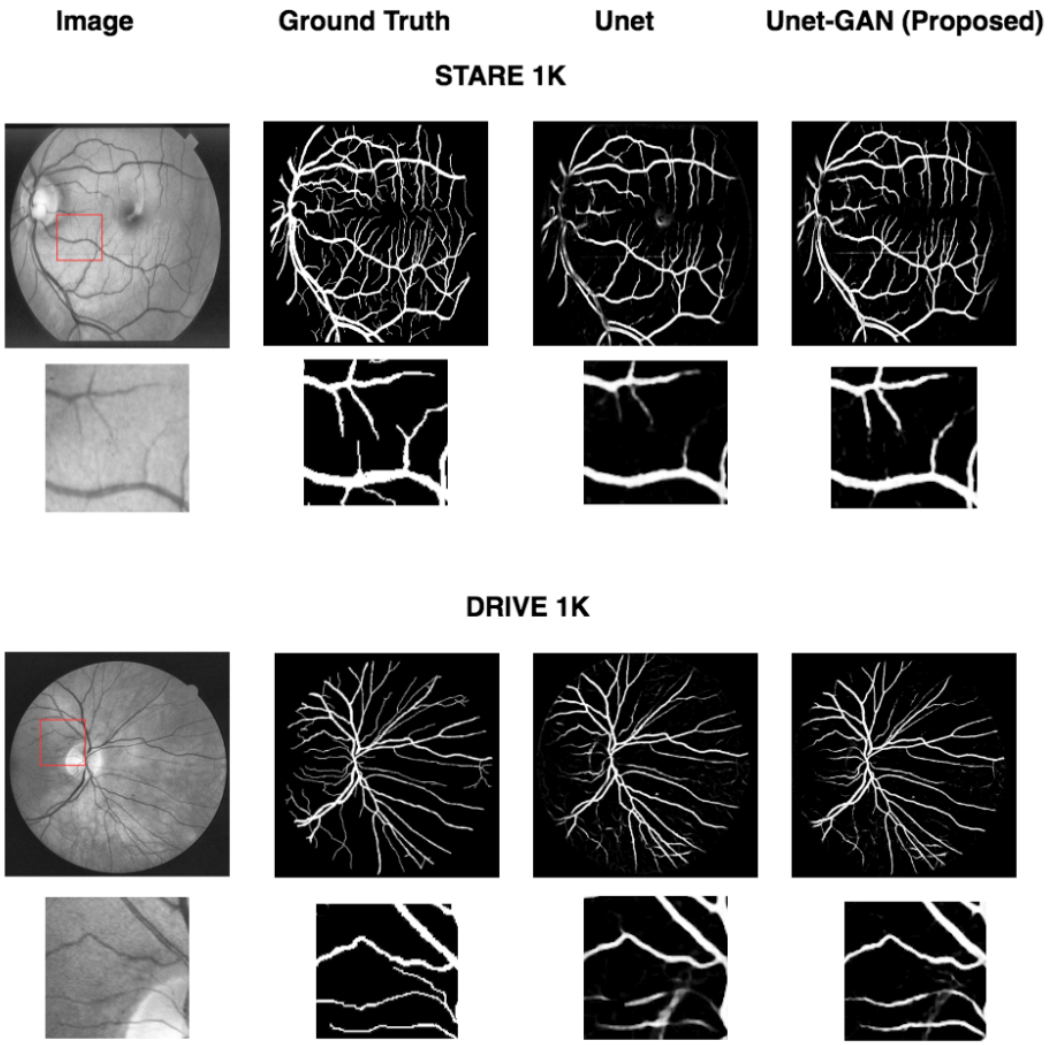


FIGURE 3.3: Some sample visualizations of segmented vessels on DRIVE and STARE dataset at 1K patch annotation budget. For each figure, we also show a zoomed up section. The effect is more pronounced on STARE dataset which consists of data from patient group with various opthalmic disorders.

3.4.1.1 Dataset

We first used data from the iSEG-2017 Challenge on infant brain MRI segmentation (62). The goal of this challenge is to compare (semi-) automatic algorithms for the segmentation of infant (~6 months) T1- and T2-weighted brain MRI scans into three tissue classes: white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF). This dataset was chosen to substantiate our proposed method: it contains the 3D multi-modal brain MRI data of only 10 labeled subjects, each one requiring about a week to annotate manually. Additionally, 13 unlabeled testing subjects are also

provided. To further validate results, we also tested our method on segmenting adult brain tissues (i.e., WM, GM, CSF) from the MRBrains-2013 (63) Challenge dataset, which contains the T1, T1-IR and FLAIR scans of 20 adult subjects. Ground truth labels are provided for only 5 training subjects, which form the training set. The test set contains the unlabeled scans of the 15 remaining subjects.

3.4.1.2 Evaluation

Segmentation accuracy is assessed using two well-known metrics, respectively measuring spatial overlap and surface distance (64):

- *Dice similarity coefficient* (DSC): This widely-used metric compares segmented volumes based on their overlap. Given a reference segmentation S_{ref}, the DSC of a predicted segmentation S_{pred} is defined as

$$\text{DSC} \;=\; \frac{2\,|S_{\text{pred}} \cap S_{\text{ref}}|}{|S_{\text{pred}}| + |S_{\text{ref}}|}. \tag{3.19}$$

  DSC values range between 0 and 1, with 1 corresponding to a perfect overlap.

- *Average Symmetric Surface Distance* (ASD): This metric computes an average of distances from points on a surface to the nearest point on another surface. Let B_{ref} and B_{pred} be the reference and predicted segmentation boundaries, it can be defined as

$$ASD \;=\; \frac{1}{N}\left(\sum_{x \in B_{\text{pred}}} d(x, B_{\text{ref}}) \;+ \sum_{x \in B_{\text{ref}}} d(x, B_{\text{pred}})\right), \tag{3.20}$$

  where $N = |B_{\text{pred}}| + |B_{\text{ref}}|$.

3.4.2 Implementation Details

3.4.2.1 Architecture

The state-of-art 3D U-Net (65) model was chosen as segmentation network in our architecture. In order to use this model in the proposed GAN framework, the following

changes were made:

- Batch-normalization (21) was replaced by weight-normalization (66), since the former had detrimental effect on GAN training for semi-supervised learning.
- As suggested in (48), ReLUs were changed to leaky ReLUs, allowing a small gradient for non-active units (i.e., units whose output is below zero).
- Max pooling was replaced by average pooling, as it leads to sparse gradient which was shown to hamper GAN training.

These modifications to 3D U-Net have helped make the training more stable and improve the performance. Other elements of the discriminator's architecture are the same as in the original U-Net.

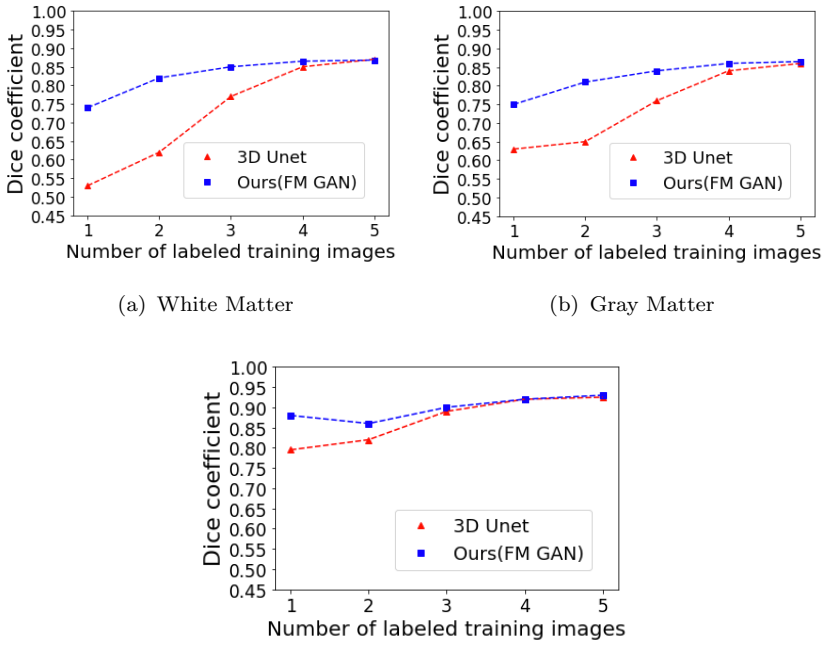
For generating 3D patches, we chose the volume generator proposed by Wu et al. (67), which was shown to provide good results for various types of 3D objects. This model leverages the power of both general-adversarial modeling and volumetric convolutional networks to generate realistic 3D shapes. For implementing the encoder, we used a standard three-layer 3D CNN architecture, whose output vector is twice the size of the generator's input noise vector. This network estimates the mean and standard deviation of the noise vector from which the given image is generated. It was found during preliminary experiments that using batch normalization in the generator and encoder gives best results. Therefore, this normalization setting was used for our GAN-based model.

3.4.2.2 Training

To train the proposed GAN based model, the 10 labeled subjects data (i.e., examples) of the iSEG-2017 dataset were split into training (1 or 2 examples), validation (1 example) and testing (7 fixed examples). The 13 unlabeled examples of the testing dataset were instead used to train the GAN.

Similarly, for the MR Brains 2013 dataset, the 5 labeled examples were split into 1 training, 1 validation and 3 testing examples, respectively. As before, the 15 unlabeled subject data were used as unlabeled data for training the GAN.

As preprocessing, N4 bias field correction was applied to images, followed by intensity normalization. To train the model, 32×32×32 patches were extracted from 3D scans with a step size of 8 voxels in each dimension. This serves two purposes: reduce computational requirements compared to employing whole 3D images, and increase the number and diversity of training examples. No other data augmentation was used, as our goal is to compare the performance of the two models in a few-shot learning scenario, not to achieve state-of-the-art performance on the tested datasets. The Adam optimizer was employed for mini-batch stochastic gradient descent (SGD), with a batch size of 30. For all networks (i.e., U-Net based discriminator, generator and encoder), we used a learning rate of 0.0001 and a momentum of 0.5.



(a) White Matter

(b) Gray Matter

(c) Cerebrospinal Fluid

FIGURE 3.4: DSC of U-Net and our FM GAN model when training with an increasing number of labeled examples from the iSEG-2017 dataset. Performance is measured on 4 test examples, and a single validation example is used during training. For our GAN-based model, the 13 unlabeled examples of the dataset are also employed during training.

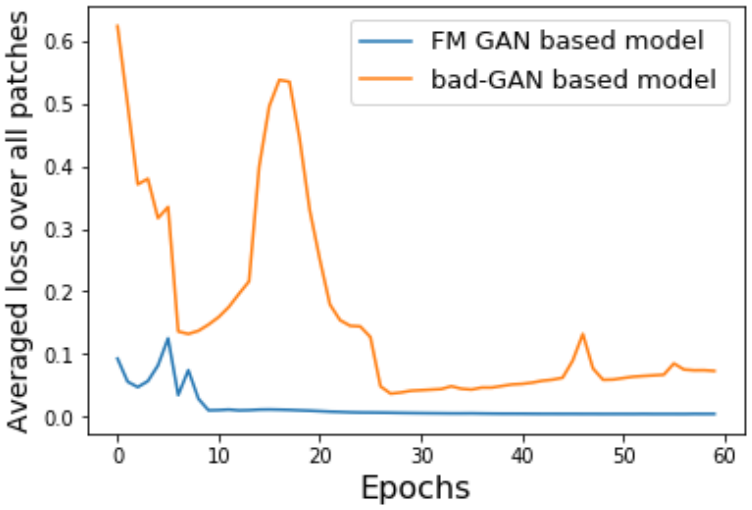


FIGURE 3.5: Feature Matching loss of bad-GAN and FM GAN models, measured at different training epochs.

3.4.3 Detailed analysis of GAN models

3.4.4 Few-shot learning with FM GAN

To validate the proposed model in a few-shot learning scenario, we trained it end-to-end with only 1 or 2 training examples. The objective is to show that, when training with few labeled examples, our model outperforms U-Net and gives performance close to full-supervision without data augmentation. While training, the model is validated with a single labeled example, thus making the total number of labeled examples no greater than 3. To reduce bias while estimating performance, we repeated this process with 3 different combinations of training and validation examples, while keeping the 7 test examples fixed, and report the average result.

Table 3.4 & 3.5 give the mean DSC and ASD obtained by the 3D U-Net modified as described in Section 3.4.2.1 (Basic U-Net), and our proposed model with standard adversarial loss (Normal GAN), feature matching (FM GAN), or the complementary GAN model of Section 3.2.3 (bad-GAN). Results are reported for 1 and 2 labeled training examples. We see that the proposed GAN-based method significantly outperforms basic U-Net when a single labeled example is available, with

TABLE 3.3: DSC and ASD (mm) results on 7 test images and 1 training image from the iSEG 2016 dataset. Best results are highlighted in bold.

Method	**WM**		**GM**		**CSF**	
	DSC	**ASD**	**DSC**	**ASD**	**DSC**	**ASD**
U-Net	0.61	1.89	0.49	2.25	0.80	0.60
Ours (normal GAN)	0.66	1.75	0.62	1.91	0.81	0.62
Ours (FM GAN)	**0.74**	**0.82**	**0.72**	**0.85**	**0.89**	**0.27**
Ours (bad-GAN)	0.69	1.20	0.68	1.33	0.86	0.39

TABLE 3.4: DSC and ASD (mm) results on 7 test images and 2 training image from the iSEG 2016 dataset. Best results are highlighted in bold.

Method	**WM**		**GM**		**CSF**	
	DSC	**ASD**	**DSC**	**ASD**	**DSC**	**ASD**
U-Net	0.68	1.43	0.62	1.79	0.82	0.59
Ours (normal-GAN)	0.71	0.96	0.72	0.89	0.82	0.51
Ours (FM GAN)	**0.80**	**0.54**	**0.80**	**0.58**	**0.88**	**0.25**
Ours (bad-GAN)	0.74	0.69	0.76	0.66	0.84	0.41

DSC improvements of 5-8% for WM, 13-23% for GM, and 1-9% for CSF. Important improvements are also observed for 2 labeled examples, with a DSC increase of 3-12%, 10-18% and 1-6% for WM, GM and CSF, respectively. Similarly, we see a significant reduction in ASD for both cases.

Comparing the different GAN models, we find that feature matching (without entropy term) yields the best performance, for all tissue classes and test cases. Compared to bad-GAN, it provides DSC improvements of 5% for WM, 4% for GM and 3% for CSF, in the case of 1 labeled example, and improvements of 6% for WM, 4% for GM and 4% for CSF, when 2 labeled examples are employed. In the next section, we analyze in greater detail the behavior of these two GAN models to better understand these results.

Next, we evaluate the impact of supervision on the performance of 3D U-Net and FM GAN by increasing the number of labeled images in training from 1 to 5. Results

of these experiments are plotted in Figure 3.4. In this experiment, we used a single validation example and a fixed set of 4 test examples.

It can be seen that, compared to U-Net, FM GAN gives a higher or equal DSC in all cases, and that the accuracy of models is comparable for 5 labeled examples. Although 5 examples seems like a relatively small number, one should remember that networks are trained using patches sampled over these images, and thus these networks see thousands of training patches.

To visually appreciate the performance of the proposed model, Figure 3.6 shows the segmentation output of Basic U-Net and FM GAN for two different subjects, when training with 1, 2 or 5 labeled examples. If 1 or 2 labeled examples are used, standard U-Net gives poor results, showing the inability of this model to work in a few shot learning scenario. In contrast, FM GAN can better learn the structure of brain tissues by using unlabeled images. Moreover, following the results of Fig. 3.4, we see that the segmentation of FM GAN is visually similar to U-Net when 5 labeled images are employed in training.

Results of the previous experiment showed the proposed model to outperform standard U-Net when very few labeled images are provided in training. In this section, we try to explain how the unlabeled and fake components of the loss function enable such improvements. Moreover, we analyze the tested GAN models to determine which elements contribute to having accurate segmentations.

Figure 3.7 plots the training losses of U-Net and our FM GAN model, at different training epochs, when using a single labeled example. For U-Net, we show the cross-entropy loss of Eq. (3.6) and validation error (i.e., mean percentage of incorrectly predicted voxels in randomly selected patches of validation images). In the case of FM GAN, we also report the unlabeled image of Eq. (3.7) and fake image loss of Eq. (3.8). These plots clearly show how U-Net, being a high-capacity model, quickly overfits the data. In contrast, our FM GAN model also learns from unlabeled and generated data and, hence, generalizes better the validation data.

To better asses the impact on segmentation of adding unlabeled and generated images, Table 3.5 gives the mean unlabeled and fake loss of the discriminator computed over test data. For this experiment, we extracted labeled patches from test images and generated an equal number of fake patches with the different GAN models. The

TABLE 3.5: Mean unlabeled and fake loss computed over patches extracted from test images, when training with 2 labeled examples. Best results highlighted in bold.

Method	**Unlabeled loss**	**Fake loss**
Basic U-Net	0.004	3.6
Ours (normal GAN)	0.0015	0.0060
Ours (FM GAN)	0.0014	**0.0020**
Ours (bad-GAN)	**0.0012**	0.0052

high fake loss value of simple U-Net confirms that this model cannot discriminate between real and fake data. This limitation of U-Net can also been seen in Fig. 3.8, which gives the predicted probabilities of U-Net and our FM GAN model for a fake input patch. Unlike U-Net, the proposed model gives a fake class probability near to 1 (i.e., white color) for all voxels of the patch.

TABLE 3.6: DSC and ASD (mm) results on 3 test images and 1 training image from the MRBrains 2013 dataset. Best results are highlighted in bold.

Method	**WM**		**GM**		**CSF**	
	DSC	**ASD**	**DSC**	**ASD**	**DSC**	**ASD**
U-Net	0.66	1.78	0.67	1.75	0.44	3.30
Ours (FM GAN)	**0.75**	**0.96**	**0.72**	**1.10**	**0.55**	**2.04**

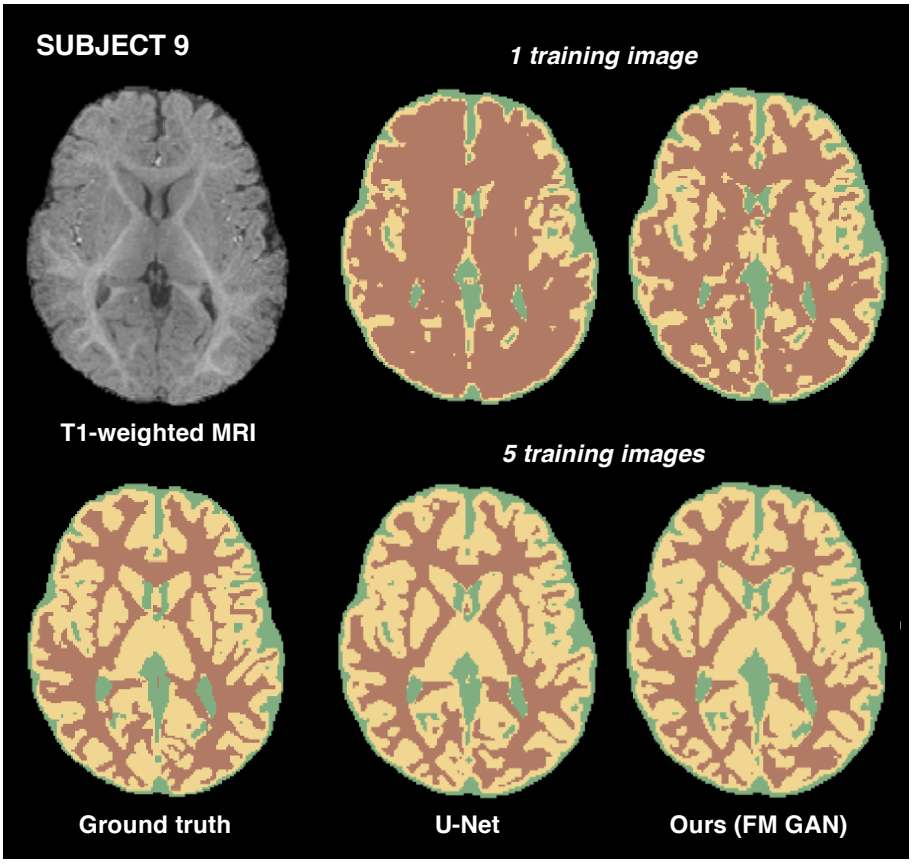
Our results indicate that FM GAN outperforms the more complex bad-GAN model (see Table 3.4 and Fig. 4.2), which also adds an entropy term to have a more diverse distribution of generated examples. For bad-GAN, we only incorporated the variational inference (VI) loss, as the low density enforcement term was not relevant in our setting given the poor sample quality. It was found that adding the VI term (46) does not improve the performance of FM GAN for semi-supervised 3D image segmentation. One possible explanation for this is the poor sample quality, which is further aggravated when increasing the entropy.

Figure 3.5 plots the feature matching loss for both the FM GAN and bad-GAN models. It can be seen that the feature matching loss of FM GAN converges quickly

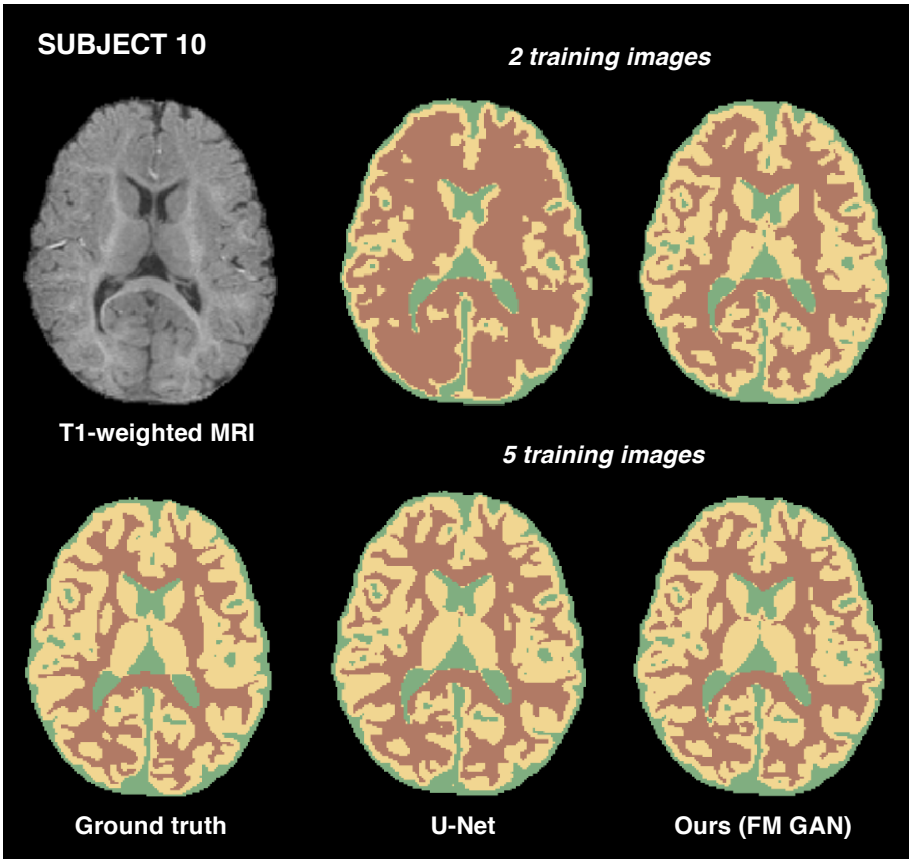
and remains less than that of bad-GAN, indicating a better sample generation. Patches generated by bad-GAN have a higher chance of being far from the true distribution and, hence, we may fail to learn a discriminator with a tight boundary of the true manifold. For example, there might be generated patches which are outside the true manifold but classified as true by the discriminator. This can also be seen in Table 3.5, where the average fake loss of bad-GAN is greater than that of FM GAN. Overall, the fake loss has an important contribution to performance in semi-supervised segmentation. It should produce samples that are different from true unlabeled images, while remaining close enough so that the discriminator learns useful information.

3.4.5 Validation on MR Brains dataset

To validate our results, we also ran similar experiments on the MR Brains dataset using just 1 training example, the results of which are listed in Table 4.4. As in previous experiments, we see that the proposed technique outperforms standard U-Net, with DSC improvements of 13.6% for WM, 7.5% for GM, and 34% for CSF. Likewise, our technique also yields a significant reduction in ASD: 46% for WM, 37% for GM, and 38% for CSF. These results suggest the usefulness of our method for across different 3D multi-modal segmentation tasks.

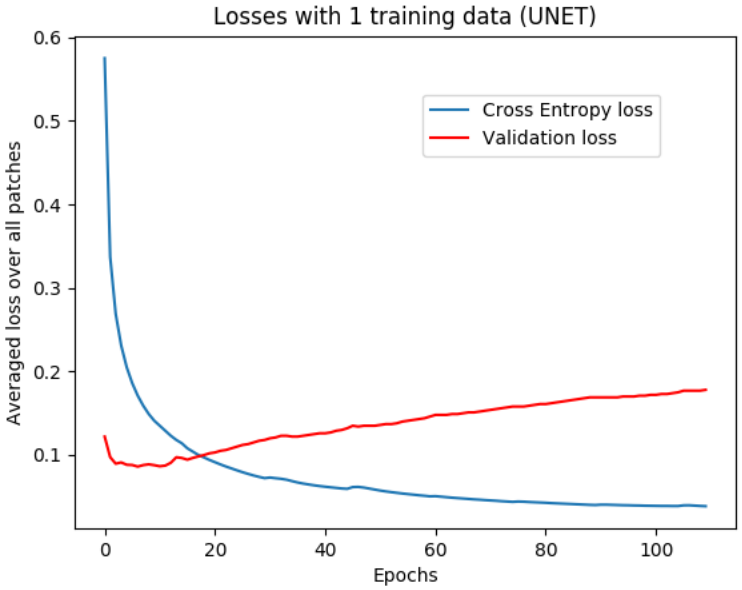


(a) Sub9

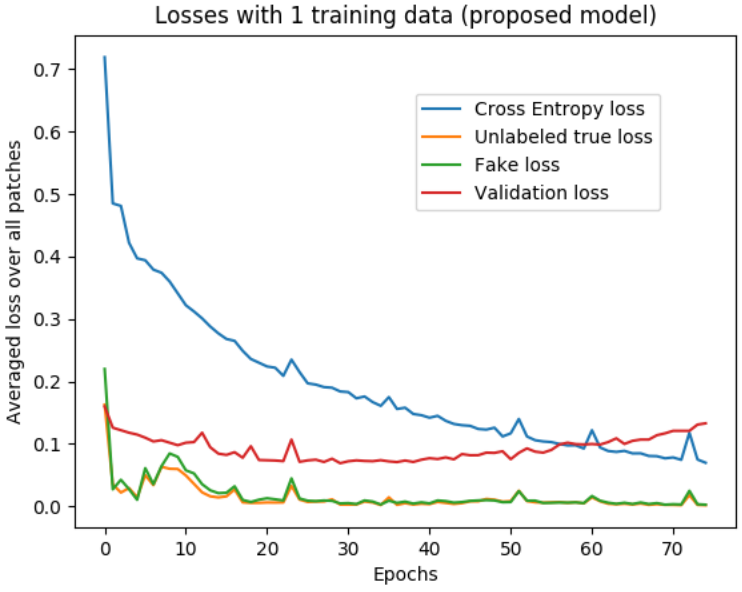


(b) Sub10

FIGURE 3.6: Visual comparison of the segmentation by each model, for two test subjects of the iSEG-2017 dataset, when training with different numbers of labeled examples.

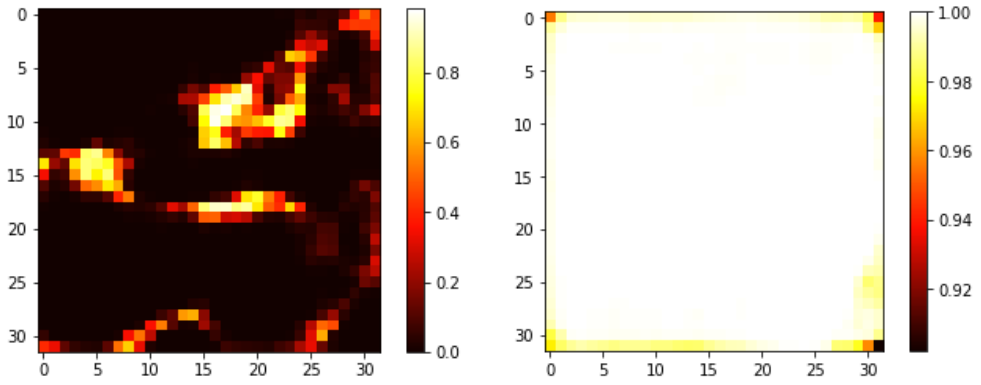


(a) U-Net



(b) Ours (FM GAN)

FIGURE 3.7: Loss of U-Net and our FM Gan model at different training epochs, measured on a random subset of validation patches.



(a) U-Net (b) Our model

FIGURE 3.8: Fake class probability predicted by U-Net and our FM GAN model for an input fake patch. Note that patches are 3D, and a single 2D slice is shown here for visualization purposes.

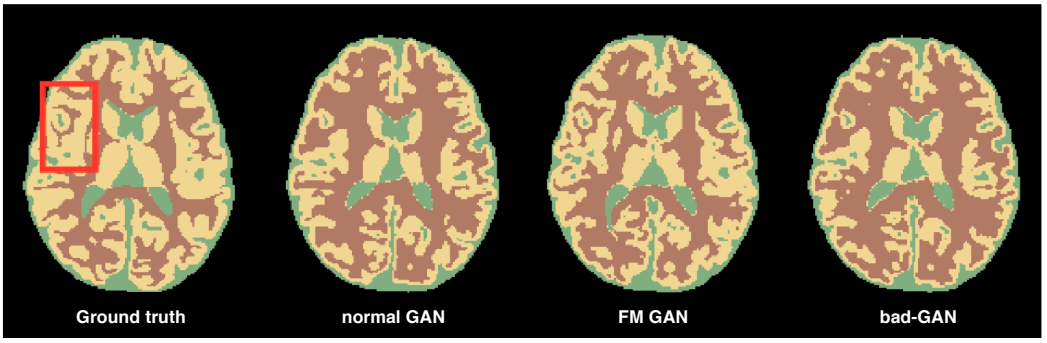


FIGURE 3.9: Segmentation of Subject 10 of the iSEG-2017 dataset predicted by different GAN-based models, when trained with 2 labeled images. The red box highlights a region in the ground truth where all these models give noticeable differences.

Chapter 4

Generator based model

4.1 Introduction

4.1.1 Generator based segmentation model

The second proposed architecture for the semi-supervised segmentation of medical images is illustrated in Figure 4.1.In this technique we place the segmenter network as a generator. We also consider a dummy network which learns to generate images from groundtruth. This idea is based on the concept of cycle-GAN (44) which is the best state of the art technique to learn unpaired image to image translation. We try learn an unpaired translation between unlabeled images and the ground truth of labeled images which provides us with the extra information about leaning how to segment the images of that dataset. We simultaneously train two generator and two discriminator network like a normal cycle GAN. The detailed theory behind this model is explained in the next section.

4.1.2 Methodology

Now we will formally write the losses and show how the model is trained in semi-supervised fashion. Before that let us present the entire architecture of the proposed model and define a few parameters. As shown in Figure 4.1 there are four major

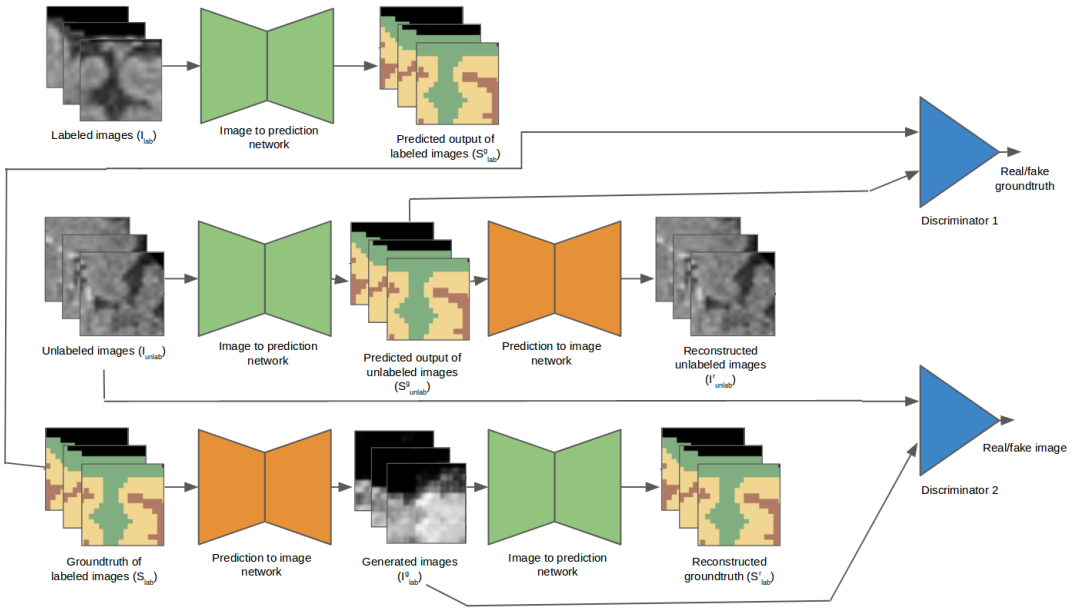


FIGURE 4.1: Schematic explaining the working of our second model. The model contains four networks which are trained simultaneously.

networks in our model: two discriminator networks- Discriminator 1(D_S) & Discriminator 2(D_I) and two generator- Image to Prediction network(G_{IS}) & Prediction to Image network(G_{SI}). In the given problem we assume that there are three distribution of images: Labeled Images (I_{Lab}), Paired Ground Truth of labeled images (S_{Lab}) and unlabeled images (I_{Unlab}). We use standard cross entropy loss to learn the corresponding output prediction of Labeled image dataset. The equation shows the cross entropy loss as a function of network G_{IS}

$$L_{CrossEntropy}(G_{IS}) = -E_{i,s\sim I_{Lab},S_{Lab}}[log(p_{GIS}(s|i))] \tag{4.1}$$

As we have another generator network G_{SI} which learns a mapping from the predictions back to the image, we also use a stardard pixelwise L1 norm between the image and the generated image to compute the loss as a function of network G_{SI}

$$L_{Norm}(G_{SI}) = E_{s,i\sim S_{Lab},I_{Lab}}[log(\|G_{SI}(s) - i\|))] \tag{4.2}$$

Both these loss component constitutes the supervised loss and now we will compute

the unsupervised loss. We first feed the G_{IS} network unlabeled image and obtain predicted out of that unlabeled image. We then pass the predicted output through G_{SI} network to reconstruct back the actual image. Meanwhile we compute an adversarial loss by passing the discriminator both predicted output and actual ground truths.

$$L_{Unlabadv}(G_{IS}, D_S) = E_{s\sim S_{Lab}}[logD_S(s)] + E_{i\sim I_{Unlab}}[log(1 - D_s(G_{IS}(i)))] \quad (4.3)$$

Finally we compute the loss in an entire cycle by taking a norm between the actual unlabelled image and reconstructed image. The expression for the same is given below:

$$L_{Cycle}(G_{IS}, G_{SI}) = E_{i\sim I_{Unlab}}[\|G_{SI}(G_{IS}(i)) - i\|] + E_{s\sim S_{Lab}}[\|G_{IS}(G_{SI}(s)) - s\|] \quad (4.4)$$

We combine all the component of losses and calculate the total loss. All the four networks are trained simultaneously. The discriminator and the generator plays a minmax game where both G_{IS} and G_{SI} try to minimize the total loss while D_S and D_I maximize it.

$$\begin{aligned} L_{Total} =& L_{Unlabadv}(G_{IS}, D_S) + L_{Unlabadv}(G_{SI}, D_I) + \gamma L_{CrossEntropy}(G_{IS}) \\ &+ \gamma L_{Norm}(G_{SI}) + \lambda L_{Cycle}(G_{IS}, G_{SI}) \end{aligned} \quad (4.5)$$

$$argmin_{G_{IS},G_{SI}} argmax_{D_S,D_I} L_{Total} \quad (4.6)$$

4.2 Experiments with 2D medical imaging dataset

We chose the same dataset which is used in the previous chapter and same network architecture for the image to prediction generator to have a fair performance. For ground truth to image generation we used a Res-Net generator. For discriminator we chose a normal 3 layered patch discriminator. Even for this experiment we took patches instead of the entire image and performed the experiments. We kept the same batch size and optimizer. We compare the results of using cycleGAN with other

TABLE 4.1: Comparison of competing supervised and semi supervised methods on DRIVE dataset.(AUC values reported)

Genre	Method	Annotated Patches			
		0.5K	1K	3K	10K
Supervised	Dasgupta *et al.*(61)	0.85	0.87	0.89	0.92
	Liskowski *et al.*(55)	0.83	0.84	0.87	0.92
	U-Net	0.89	0.90	0.92	0.95
Semi Supervised	Lahiri *et al.* (49)	0.82	0.84	0.85	0.93
	Proposed model(FM GAN)	0.92	0.94	0.96	0.965
	Proposed model(bad GAN)	0.926	0.945	0.96	0.965
	Proposed model(cycle GAN)	0.935	0.95	0.965	0.97

state of the art techniques and the proposed idea which is provided in the previous chapter. It is clear from both the tables that for 2D medical image segmentation the cycleGAN technique provides the best results after bad-GAN and FM-GAN based methods. In the next section we validate something similar for the 3D medical image segmentation task.

TABLE 4.2: Comparison of competing supervised and semi supervised methods on STARE dataset.(AUC values reported)

Genre	Method	Annotated Patches			
		0.5K	1K	3K	10K
Supervised	Dasgupta *et al.*(61)	0.82	0.84	0.87	0.91
	Liskowski *et al.*(55)	0.84	0.86	0.89	0.93
	U-Net	0.86	0.89	0.90	0.94
Semi Supervised	Lahiri *et al.* (49)	0.80	0.81	0.83	0.90
	Proposed model(FM GAN)	0.90	0.92	0.94	0.96
	Proposed model(bad GAN)	0.91	0.923	0.943	0.96
	Proposed model(cycle GAN)	0.924	0.94	0.95	0.965

4.3 Experiments with 3D medical imaging dataset

Again for this part we took the same dataset and run the experiments to test our new model. We used 3D U-Net for G_{IS} and 3D Res-Net for G_{SI}. Also, for the discriminator we used a standard 3D discrimnator. We report our results in the following tables:

TABLE 4.3: DSC and ASD (mm) results on 7 test images and 1 training image from the iSEG 2016 dataset. Best results are highlighted in bold.

Method	**WM**		**GM**		**CSF**	
	DSC	**ASD**	**DSC**	**ASD**	**DSC**	**ASD**
U-Net	0.61	1.89	0.49	2.25	0.80	0.60
Ours (normal GAN)	0.71	0.96	0.72	0.89	0.82	0.51
Ours (FM GAN)	0.74	0.82	0.72	0.85	0.89	0.27
Ours (bad-GAN)	0.69	1.20	0.68	1.33	0.86	0.39
Ours (cycle GAN)	**0.79**	**0.76**	**0.78**	**0.75**	**0.92**	**0.16**

TABLE 4.4: DSC and ASD (mm) results on 3 test images and 1 training image from the MRBrains 2013 dataset. Best results are highlighted in bold.

Method	**WM**		**GM**		**CSF**	
	DSC	**ASD**	**DSC**	**ASD**	**DSC**	**ASD**
U-Net	0.66	1.78	0.67	1.75	0.44	3.30
Ours (FM GAN)	0.75	0.96	0.72	1.10	0.55	2.04
Ours (cycle GAN)	**0.79**	**0.86**	**0.76**	**0.95**	**0.61**	**1.60**

For iSEG dataset we show comparison between all the proposed techniques and their performances. Below we provide an image for visual comparison of a subject between all the techniques for iSEG dataset.

4.4 Conclusion and Future Work

We presented multiple methods for segmenting both 2D and 3D multi-modal medical images, which can achieve performances comparable to full-supervision with only a few training samples using semi-supervised learning. In both the techniques, we showed how the method uses unlabeled data to prevent over-fitting and learn in a semi-supervised manner, by learning to discriminate between true and generated fake patches in first case and learning an extra unpaired translation between unlabeled images and labeled ground truths. The proposed models can be employed to

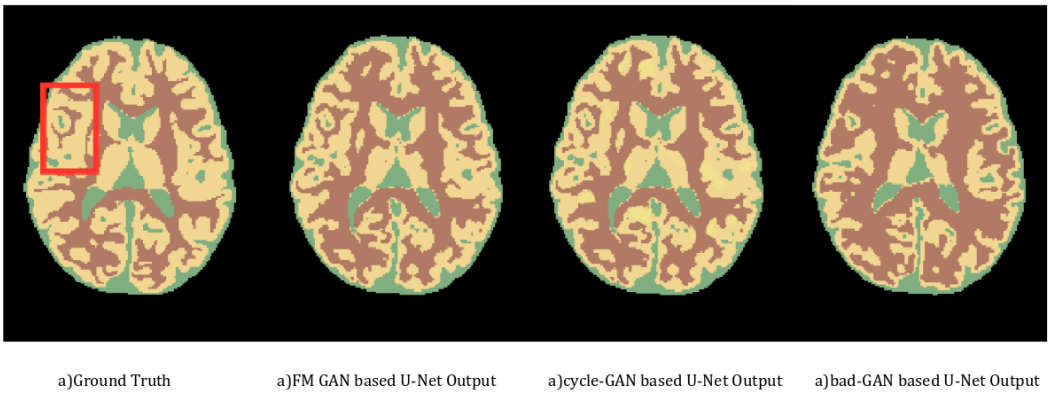


FIGURE 4.2: Segmentation of Subject 10 of the iSEG-2017 dataset predicted by different GAN-based models, when trained with 2 labeled images. The red box highlights a region in the ground truth where all these models give noticeable differences.

enhance any segmentation network in a low data setting, where the network fails to produce a good segmentation output. It also provides a new technique for few-shot learning, obviating the need for an initial pre-trained network by leveraging the semi-supervised learning ability of GANs. Moreover, results on the DRIVE, STARE, iSEG-2017 and MRBrains 2013 datasets showed our method's potential for reducing the burden of acquiring annotated medical data or any segmentation data in general. Our experiments explored different GAN based architectures and their impact on segmentation performance. We showed empirically that cycle GAN based model performs better than the other GAN based models for segmenting images in a semi-supervised fashion. This work can be further extended in future to design new techniques for semi-supervised semantic segmentation. The cycleGAN based technique can be explored further theoretically which will give us more insights on how it is the most efficient algorithm.

Bibliography

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[3] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[4] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4507–4515.

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[6] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[7] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, and I. Išgum, "Automatic segmentation of mr brain images with a convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1252–1261, 2016.

[8] P. Naylor, M. Laé, F. Reyal, and T. Walter, "Nuclei segmentation in histopathology images using deep neural networks," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 933–936.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[10] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE transactions on medical imaging*, vol. 35, no. 11, pp. 2369–2380, 2016.

[11] L. Lin, W. Yang, C. Li, J. Tang, and X. Cao, "Inference with collaborative model for interactive tumor segmentation in medical image sequences," *IEEE transactions on cybernetics*, vol. 46, no. 12, pp. 2796–2809, 2016.

[12] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "Endonet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging*, vol. 36, no. 1, pp. 86–97, 2017.

[13] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, 2017.

[14] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. C. Mok, L. Shi, and P.-A. Heng, "Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1182–1195, 2016.

[15] M. J. van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1273–1284, 2016.

[16] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5 d representation for lymph node detection

using random sets of deep convolutional neural network observations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 520–527.

[17] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

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

[19] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on psychological science*, vol. 6, no. 1, pp. 3–5, 2011.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[22] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[23] P. O. Pinheiro and R. Collobert, "Weakly supervised semantic segmentation with convolutional networks," in *CVPR*, vol. 2, no. 5. Citeseer, 2015, p. 6.

[24] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep CNN for semantic image segmentation," *arXiv preprint arXiv:1502.02734*, 2015.

[25] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. Ben Ayed, "Constrained-cnn losses forweakly supervised segmentation," *arXiv preprint arXiv:1805.04628*, 2018.

[26] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1796–1804.

[27] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1635–1643.

[28] M. Rajchl, M. C. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz *et al.*, "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 674–683, 2017.

[29] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3159–3167.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[31] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5689–5697.

[32] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[33] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert, "Semi-supervised learning for network-based cardiac mr image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 253–260.

[34] C. Baur, S. Albarqouni, and N. Navab, "Semi-supervised deep learning for fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 311–319.

[35] S. Min and X. Chen, "A robust deep attention network to noisy labels in semi-supervised biomedical segmentation," *arXiv preprint arXiv:1807.11719*, 2018.

[36] Y. Zhou, Y. Wang, P. Tang, W. Shen, E. K. Fishman, and A. L. Yuille, "Semi-supervised multi-organ segmentation via multi-planar co-training," *arXiv preprint arXiv:1804.02586*, 2018.

[37] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2827–2836.

[38] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," *arXiv preprint arXiv:1712.04440*, 2017.

[39] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 408–416.

[40] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *arXiv preprint arXiv:1611.08408*, 2016.

[41] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[42] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

[45] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3d-guided cycle consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 117–126.

[46] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov, "Good semi-supervised learning that requires a bad gan," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6510–6520. [Online]. Available: http://papers.nips.cc/paper/7229-good-semi-supervised-learning-that-requires-a-bad-gan.pdf

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[48] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2234–2242. [Online]. Available: http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.pdf

[49] A. Lahiri, K. Ayush, P. K. Biswas, and P. Mitra, "Generative adversarial learning for reducing manual annotation in semantic segmentation on large scale miscroscopy images: Automated vessel segmentation in retinal fundus image as test case," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 42–48.

[50] Z. Dai, A. Almahairi, P. Bachman, E. Hovy, and A. Courville, "Calibrating energy-based generative adversarial networks," *arXiv preprint arXiv:1702.01691*, 2017.

[51] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *ICLR*, 2016.

[52] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[54] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE transactions on medical imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[55] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE transactions on medical imaging*, vol. 35, no. 11, pp. 2369–2380, 2016.

[56] A. Lahiri, A. G. Roy, D. Sheet, and P. K. Biswas, "Deep neural ensemble for retinal vessel segmentation in fundus images towards achieving label-free angiography," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the.* IEEE, 2016, pp. 1340–1343.

[57] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[58] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis." in *CVPR*, vol. 1, no. 2, 2017, p. 3.

[59] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 901–909.

[60] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[61] A. Dasgupta and S. Singh, "A fully convolutional neural network based structured prediction approach towards the retinal vessel segmentation," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 248–251.

[62] L. Wang, D. Nie, G. Li, Puybareau, J. Dolz, Q. Zhang, F. Wang, J. Xia, Z. Wu, J. Chen, K. Thung, T. D. Bui, J. Shin, G. Zeng, G. Zheng, V. S. Fonov, A. Doyle, Y. Xu, P. Moeskops, J. P. W. Pluim, C. Desrosiers, I. Ben Ayed, G. Sanroma, O. M. Benkarim, A. Casamitjana, V. Vilaplana, W. Lin, G. Li, and D. Shen, "Benchmark on automatic 6-month-old infant brain segmentation algorithms: The iseg-2017 challenge," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2019.

[63] A. M. Mendrik, K. L. Vincken, H. J. Kuijf, M. Breeuwer, W. H. Bouvy, J. De Bresser, A. Alansary, M. De Bruijne, A. Carass, A. El-Baz *et al.*, "Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans," *Computational intelligence and neuroscience*, vol. 2015, p. 1, 2015.

[64] V. Yeghiazaryan and I. Voiculescu, "An overview of current evaluation methods used in medical image segmentation," Tech. Rep. CS-RR-15-08, Department of Computer Science, University of Oxford, Oxford, UK, Tech. Rep., 2015.

[65] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432.

[66] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 901–909.

[67] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," in *Advances in Neural Information Processing Systems*, 2016, pp. 82–90.