# Semi-Supervised Semantic Segmentation of Vessel Images using Leaking Perturbations

Jinyong Hou, Xuejie Ding, Jeremiah D. Deng
Department of Information Science, University of Otago
robert.hou@postgrad.otago.ac.nz, {emily.ding, jeremiah.deng}@otago.ac.nz

## Abstract

*Semantic segmentation based on deep learning methods can attain appealing accuracy provided large amounts of annotated samples. However, it remains a challenging task when only limited labelled data are available, which is especially common in medical imaging. In this paper, we propose to use Leaking GAN, a GAN-based semi-supervised architecture for retina vessel semantic segmentation. Our key idea is to pollute the discriminator by leaking information from the generator. This leads to more moderate generations that benefit the training of GAN. As a result, the unlabelled examples can be better utilized to boost the learning of the discriminator, which eventually leads to stronger classification performance. In addition, to overcome the variations in medical images, the mean-teacher mechanism is utilized as an auxiliary regularization of the discriminator. Further, we modify the focal loss to fit it as the consistency objective for mean-teacher regularizer. Extensive experiments demonstrate that the Leaking GAN framework achieves competitive performance compared to the state-of-the-art methods when evaluated on benchmark datasets including DRIVE, STARE and CHASE_DB1, using as few as 8 labelled images in the semi-supervised setting. It also outperforms existing algorithms on cross-domain segmentation tasks.*

## 1. Introduction

Semantic segmentation on medical imaging has been an important and challenging visual recognition task aims at assigning a label from a set of classes to each pixel of the image [6, 7, 26]. The segmentation techniques based on deep neural networks have been applied in various medical imaging modalities and attained the promising improvements. For fully supervised deep learning models, they rely on large volume of annotated datesets. However, for medical images, it is generally time-consuming and laborious for domain experts to label them especially for pixel-level.

To overcome this limitation, semi-supervised learning methods are explored for medical image segmentation [45, 46, 3]. But in real world, medical images are often taken from different scanners and environments, such as lights, angle, which alter the appearance of the images. Also, the images are related to the different categories of subjects (e.g. healthy or pathological), which varies the morphology of images. Both of them cause the domain shift for the different medical images. Towards these practical issues, we propose a model with generalization capacity which can transfer the knowledge from one domain to other unseen domains for the segmentation.

Specifically, a GAN-based semi-supervised strategy [34] is adopted as our basic framework. To improve its generalization, we add three improved strategies on the GAN-based framework. First, a U-Net style network is used as the discriminator due to its competitive performance in medical image segmentation [33]. Secondly, we propose a polluted discriminator by introducing auxiliary "leaking links" from the generator to the discriminator, which force the generator to produce moderate (unrealistic) generations to boost the semi-supervised learning performance. The polluted discriminator gains an improved learning ability due to the perturbation of probabilities during gradient optimization. Finally, the discriminator is regularized by the mean-teacher mechanism [41], which can improve the segmentation generalization because it is trained under input and weights perturbations. We also introduce a novel focal consistency penalty for the predictions of the student and teacher modules. It can extend the training range when the two prediction results have significant shift.

Among various medical image modalities, retinal vessel images include rich information about pathological changes that are critical for early diagnosis and treatment of eye-related diseases [44]. Also, blood vessel analysis plays a fundamental role in different clinical fields [27, 36, 4]. However, compared with other objects, the segmentation of vessels requires more fine-grained results, which is rather challenging [9]. There are variations in terms of their sizes, shapes, and intensities in different local areas, forcing the segmentation model to learn the vessel's complex structure.

Also, a micro-vessel can be as thin as just one to several pixels. On the other hand, there are a number of public-domain retinal vessels datasets, such as DRIVE [40], STARE [17], and CHASE_DB1 [10]. This provides a practical platform for us to utilize different datasets to evaluate our model's transfer learning [29] performance.

## 2. Related Work

**Vessel Segmentation based on Deep Neural Networks:** Blood vessel segmentation is a hot topic in medical image analysis since the analysis results play an essential role in the diagnosis and intervention of many diseases [36]. For retinal vessel images, it was pointed out in [37] that the challenge in vessel segmentation is due to the unique complex structure of blood vessels. Deep convolutional neural networks applied to vessel segmentation have obtained some competitive results, and they allow a direct end-to-end function that can be trained using back propagation [11, 44, 22, 43, 19, 30]. Also, some models [23, 10, 14] can be used for cross-training evaluations, i.e., training on one dataset and testing on another dataset. It helps the framework be deployed to test fundus images obtained with different cameras and different environments. These successes, however, rely on ample annotated data, which usually is not the case for medical images as labelling full images is time-consuming and expensive. The limited availability of annotated vessels remains the bottleneck for deep neural networks' application in real-world vessel segmentation tasks.

**Semi-Supervised Segmentation:** To address the data limitation, several semi-supervised semantic segmentation methods have been proposed for medical images [46, 3, 45]. The unlabelled data can be used freely to improve fully supervised performance. The generative adversarial network (GAN) has shown potential performance on semi-supervised learning [16, 34, 12]. The discriminator also serves as a classifier, not only discriminates real examples from fake ones but also classifies the former [34]. Using the GAN, the generator can create extensive realistic visual data, enabling the discriminator to learn better features for more accurate pixel classification in semantic segmentation [38]. It has been pointed out in [8, 20] that good semi-supervised classification performance and perfect generators could not be obtained simultaneously in the GAN-based methods. Further insights were provided in [20] on how fake examples influence semi-supervised learning, and that only moderate fake examples can its boost performance. Inspired by these works, we propose a novel GAN method that can generate moderate samples and force the discriminator to better learn the fine-grained vessel structure.

Another semi-supervised state-of-the-art technique is Mean-Teacher, a method that averages model weights to form a target-generating teacher model [41, 2]. The mean-teacher paradigm is applied in cross-domain detection [5],

where the domain gap was bridged with consistency regularization in a teacher-student scheme. In our work, the mean-teacher is explored to improve the segmentation generalization capacity by the perturbation from inputs and weights in a teacher-student scheme. Meanwhile, we also propose a focal consistency to penalize the teacher and student networks' predictions. The novel consistency makes the network focus on the uncertain pixels, especially for the thin vessel, challenging to distinguish.

**Skip Connection in Deep Learning:** The use of skip connections often takes a crucial role in deep neural networks classifiers [15, 21, 18, 39] as it can effectively mitigate the vanishing gradient problem. Another research line combines the shallow information from lower layers with deep information from high layers [33, 25, 32] and hence leads to more accurate recognition results. In [35], the authors propagate the discriminator information back to the generator, utilizing an iterative feedback loop to improve the generator's performance. Reversely, in our work, we employ skip connections to 'leak' information from the generator to the discriminator to boost the performance of semi-supervised learning.

## 3. Proposed Model

### 3.1. Problem setting

We have a training set $\mathcal{D}_s$ that contains labelled samples $\{(\boldsymbol{x}_l^i, y^i)|i = 1, \cdots, n\}$ and unlabelled samples $\{\boldsymbol{x}_{ul}\}$, where each sample $\boldsymbol{x}_l, \boldsymbol{x}_{ul} \in \mathcal{X}$, only limited samples $\boldsymbol{x}_l^i$ are associated with a label $y_i \in \mathcal{Y}$. $\mathcal{X}$ is the feature space, such as an $m$ dimensional space of real numbers $\mathbb{R}^m$, and $\mathcal{Y}$ is the label space. The unlabelled data $\boldsymbol{x}_{ul}$ is used to improve the predictions of the classifier. The classifier can provide labels for unlabelled data.

Our goal is to design a generalization segmentation framework to achieve good performance on semi-supervised learning scenarios. It can also perform well in cross-training scenarios, i.e., training on dataset $\mathcal{D}_s$ and testing on dataset $\mathcal{D}_t$. Here the $\boldsymbol{x}_t, \boldsymbol{x}_l, \boldsymbol{x}_{ul} \in \mathcal{X}$ are in the same feature space. However, there is domain shift between $\mathcal{D}_s$ and $\mathcal{D}_t$, and therefore different distributions $p(\mathbf{X}_s) \neq p(\mathbf{X}_t)$. In the real world, the distribution discrepancy between training data and testing data will often hinder performance. Therefore, generalization segmentation networks are designed to be transferred to other domains (even unseen scenarios), and it performs well for domain shift.

### 3.2. Model framework

The proposed framework is shown in Fig. 1. Our model follows the GAN-based semi-supervised approach. We aim to classify each pixel of the retinal image to one of the $K$ classes.

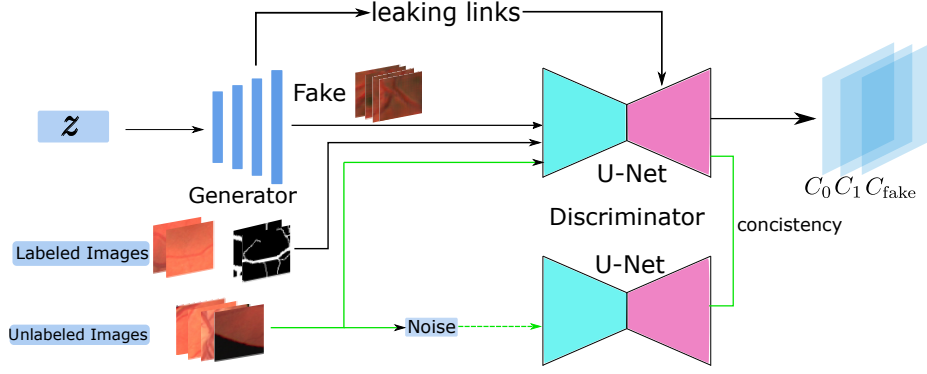In the model, the fake image $\boldsymbol{x}_f$ is labelled with an addi-

Figure 1: Model Schematic for retina semantic segmentation. There are three inputs (labelled, unlabelled and fake data) to feed into the discriminator. U-Net is utilized as the discriminator in the proposed model. During training, unlabelled data are also used to train a teacher discriminator.

tional $K + 1$ class, and $p_k(y = K + 1|\boldsymbol{x}_f)$ can be used to supply the probability that data $\boldsymbol{x}_f$ is fake. During the training, the unlabelled data $\boldsymbol{x}_{\mathrm{ul}}$ has to be one of the $K$ classes. The generated fake samples can be used with unlabelled data to explore the hidden data structure to support semantic segmentation. The labelled data $\boldsymbol{x}_l$ is utilized for training the discriminator to recognize the pixel's categories. Therefore, the discriminator's loss function is given as a sum of supervised and unsupervised parts:

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \mathcal{L}_{\mathrm{unsup}}. \tag{1}$$

The loss for supervised images is a cross-entropy loss between labels and probabilities predicted by the discriminator, as used in standard segmentation networks:

$$\mathcal{L}_{\mathrm{sup}} = -\mathbb{E}_{\boldsymbol{x}_l, y \sim p_{\mathrm{data}}(\boldsymbol{x}_l, y)} \log p_k(y|\boldsymbol{x}_l, y < K + 1). \tag{2}$$

The loss for unsupervised images is defined as

$$\mathcal{L}_{\mathrm{unsup}} = -\mathbb{E}_{\boldsymbol{x}_{\mathrm{ul}} \sim p_{\mathrm{data}}(\boldsymbol{x}_{\mathrm{ul}})} \log D(\boldsymbol{x}_{\mathrm{ul}}) \\ - \mathbb{E}_{\boldsymbol{z} \sim G} \log(1 - D(G(\boldsymbol{z})), \tag{3}$$

where $\boldsymbol{z}$ is the noise used to generate the fake images. Due to the over-parameterization of the $K + 1$ outputs in the discriminator, the $(K + 1)$-th logit, $l_{K+1}(\boldsymbol{x}_f)$ can be clamped to 0 for all the $\boldsymbol{x}_f$ [34]. Then the discriminator $D$ can be obtained by $D(\cdot) = \frac{Z(\cdot)}{Z(\cdot)+1}$ for unlabelled and fake images, where $Z(\cdot) = \sum_{i=1}^{K} \exp(l_i(\cdot))$. Also, the feature matching loss has a better performance to train the model.

In the model, the *discriminator* also functions as a classifier for semantic segmentation. It is crucial for the segmentation performance. Among various discriminator models that can be chosen for the task, we choose *U-Net* as it is a good performer on medical image segmentation.

To boost the semi-supervised learning, we propose a new leaking mechanism for the discriminator, as shown in Fig. 2.

We add a skip connection between the generator and the discriminator. Let us denote the $l$-th layer outputs of the generator, the contracting path and expanding path of the U-Net by $\{G^l\}_{l=1}^{n-1}, \{U_c^l\}_{l=n-1}^{1}$ and $\{U_e^l\}_{l=n+1}^{2n-1}$ respectively, where $n$ is the index number of the bottom layer in U-Net. The activation of the generator and the U-Net contracting path are concatenated by the *information leaking* module. We denote the $l$-th output layer of information leaking module as $\{L^l\}_{l=1}^{n}$. The activation of each layer in the information leaking module can be described as

$$L^l = [\alpha^l G^l, \beta^l U_c^l], \tag{4}$$

where $[\cdot, \cdot]$ denotes the concatenated operation. The coefficients $\alpha^l$ and $\beta^l$ are scaling parameters for the outputs in the layer. Then the output of the information leaking module is added to the corresponding activation of the expanding path, as follows:

$$U_e^l = f[U_e^l, \gamma L^l], \tag{5}$$

where $f$ is the activation function, and $\gamma$ is a switch function. When $\gamma = (1, 1)$ the information leaking module is activated; when $\gamma = (0, 1)$, the module is disabled. Thus, the information leaking module can be flexibly switched on and off.
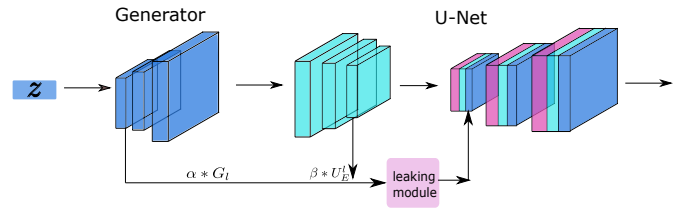


Figure 2: For the information leaking module, the outputs of the generator layers are concatenated with the upsampling parts of U-Net.

Then the discriminator gives the decision according to the features polluted by leaking information from the generator.

It leads to generate unrealistic examples with a perfect generator. The unrealistic examples can boost the semi-supervised learning performance, as proven in [20, 8]. However, different from bad GAN [8], our generator is perfect. The unrealistic examples are generated due to the polluted discriminator. Also, the polluted features induce the perturbation on the probability during the gradient optimization. These kinds of perturbations also improve the learning ability of discriminator (U-Nets). A detailed explanation of the leaking module is in section 3.3.

With the information leaking strategy, the proposed model tends to leverage the unlabelled data to explore the more intrinsic properties of data and achieve better results. However, variations factors such as image intensity changes, noise, and low contrast, still present significant obstacles for the segmentation model [27].

To further improve the discriminator's generalization over the variations in medical images, we adopt the mean-teacher paradigm [41, 5] as an additional regularization. Specifically, follow the protocol of [5], we only apply the noise to the teacher path. Also, under the medical image scenarios, we apply the added noise perturbations but not the augmentation for this. The student's inputs $\boldsymbol{x}_{\mathrm{ul}}^{s}$ polluted by noise are fed into the teacher network. That is, for the teacher's network, inputs $\boldsymbol{x}_{\mathrm{ul}}^{t}$ become $\boldsymbol{x}_{\mathrm{ul}}^{t} = \boldsymbol{x}_{\mathrm{ul}}^{s} + \lambda\epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$, $\lambda < 1$. Another perturbation is from the weights. Here we notify the weights of student discriminator $w_{D_s}$. With the same architecture as the students, the teacher discriminator is parameterized by $w_{D_t}$. According to the mean-teacher strategy [41], the teacher network weights $w_{D_t}$ are the exponential moving average (EMA) of the student weights:

$$\omega_{D_t}^{\tau} = \alpha\omega_{D_t}^{\tau-1} + (1 - \alpha)\omega_{D_s}^{\tau}, \qquad (6)$$

where the $\alpha$ is the smoothing coefficient parameter that controls the updating of teacher weights, which is usually set between 0.9 and 0.999, and the $\tau$ is the training step. Then a consistency loss [41] is employed to penalize the difference between the student's prediction $f_s(\boldsymbol{x}_{ul}^{s}; \omega_{D_s})$ and the teacher's $f_t(\boldsymbol{x}_{ul}^{t}; \omega_{D_t})$:

$$\mathcal{L}_{\mathrm{cons}}(\boldsymbol{x}_{\mathrm{ul}}) = ||f_s(\boldsymbol{x}_{ul}^{s}; \omega_{D_s}) - f_t(\boldsymbol{x}_{ul}^{t}; \omega_{D_t})||. \qquad (7)$$

### 3.3. Probability perturbation analysis

For the GAN-based semi-supervised approach, it has been pointed out that a moderate generation can boost the performance [20][8]. To obtain a moderate generation, a complementary generator (i.e., a bad generator) was introduced in [8] to improve the generalization of a semi-supervised learning framework. In our method, we keep the perfect generator to capture the data distribution but introduce a "polluted" discriminator for moderate generations.

**Definition 1** (Polluted Discriminator). The polluted discriminator distinguishes the noise-added example $\tilde{\boldsymbol{x}}$ from the gen-

erated example $\boldsymbol{x}_g$. The noise is leaked from the generator by the information leaking module. Then for any optimal solution $V(G, D)$, $p_G \approx p_d$.

$$\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\mathrm{data}}}[\log D(\tilde{\boldsymbol{x}})] + \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \qquad (8)$$

The polluted discriminator leads the generator to a distorted version of the real example; it pushes the generator away from producing strong generations. Furthermore, the distance between the distorted image $\tilde{\boldsymbol{x}}_{\mathrm{ul}}$ and the real one $\boldsymbol{x}_{\mathrm{ul}}$ can be controlled by the leaking information module. That means we can control how much the distortion the generations have.

Following the approach in [20], we now give an analysis on the probability perturbation effects of information leaking and illustrate why it works. The unsupervised objective in Eq. (3) can be rewritten as follows:

$$\mathcal{L}_{\mathrm{unsup}} = -\mathbb{E}_{\boldsymbol{x}_f \sim p_{\mathrm{fake}}} \log \frac{\exp(l_{K+1}(\tilde{\boldsymbol{x}}_f))}{\sum_{i=1}^{K+1} \exp(l_i(\tilde{\boldsymbol{x}}_f))}$$
$$- \mathbb{E}_{\boldsymbol{x}_{\mathrm{ul}} \sim p_{\mathrm{data}}} \left[ \log(1 - \frac{\exp(l_{K+1}(\tilde{\boldsymbol{x}}_{\mathrm{ul}}))}{\sum_{i=1}^{K+1} \exp(l_i(\tilde{\boldsymbol{x}}_{\mathrm{ul}}))}) \right], \qquad (9)$$

where $p_{\mathrm{model}}(y = K + 1|\boldsymbol{x})$ is replaced by $\frac{\exp(l_{K+1}(\boldsymbol{x}))}{\sum_{i=1}^{K+1} \exp(l_i(\boldsymbol{x}))}$. When we take the derivative w.r.t the weights of the discriminator, we get

$$\nabla\mathcal{L}_{\mathrm{unsup}} = \mathbb{E}_{\boldsymbol{x}_f \sim p_{\mathrm{fake}}} \sum_{i=1}^{K} \underbrace{p_{\mathrm{model}}(y = i|\boldsymbol{x}_f)\nabla l_i(\boldsymbol{x}_f)}_{a_i(\tilde{\boldsymbol{x}}_f)}$$
$$- \mathbb{E}_{\tilde{\boldsymbol{x}}_{\mathrm{ul}} \sim p_{\mathrm{data}}} \sum_{i=1}^{K} \underbrace{[p_{\mathrm{model}}(y = i|\tilde{\boldsymbol{x}}_{\mathrm{ul}})p_{\mathrm{model}}(y = K + 1|\tilde{\boldsymbol{x}}_{\mathrm{ul}})]\nabla l_i(\tilde{\boldsymbol{x}}_{\mathrm{ul}})}_{b_i(\tilde{\boldsymbol{x}}_{\mathrm{ul}})}. \qquad (10)$$

Our model leaks the information from the generator to the discriminator, concatenating as a feature vector to the expanding path. This makes the discriminator give the decision on $\tilde{\boldsymbol{x}}_{\mathrm{ul}}$ instead of the real input $\boldsymbol{x}_{\mathrm{ul}}$. Under this situation, we assume there is some perturbation on the probability in the $b_i(\tilde{\boldsymbol{x}}_{\mathrm{ul}})$ part, with the probability of being classified as normal classes reduced, and the probability of being classified as "fake" increased, all due to the pollution resulted from the leaking link. It becomes

$$b_i(\tilde{\boldsymbol{x}}_{\mathrm{ul}}) = -\mathbb{E}_{\boldsymbol{x}_{\mathrm{ul}} \sim p_{\mathrm{data}}} \sum_{i=1}^{K} [(p_{\mathrm{model}}(y = i|\boldsymbol{x}_{\mathrm{ul}}) - \varepsilon_1)$$
$$(p_{\mathrm{model}}(y = K + 1|\boldsymbol{x}_{\mathrm{ul}}) + \varepsilon_2)]\nabla l_i(\boldsymbol{x}_{\mathrm{ul}}). \qquad (11)$$

The implication of Eq. (11) can be analysed as follows.

1. The perturbation can increase the gradient at the early training. At the beginning of training, the examples generated from the generator are very weak and easy to distinguish from the real examples. There exists

$i \leq K$, $p(y = i|\boldsymbol{x}_{\text{ul}}) \approx 1$, $p(y = K + 1|\boldsymbol{x}_{\text{ul}}) \approx 0$, and $p(y = i|x_f) \approx 0$. Under this situation, there is no gradient flow from Eq. (11), i.e., zero contribution for unsupervised part in the overall gradient. This issue is remedied when the leaking information is added to pollute the unlabelled feature. As a result, the $p(y = K + 1|\boldsymbol{x}_{\text{ul}}) \approx 0 + \varepsilon_2$ and $p(y = i|\boldsymbol{x}_f) \approx 1 - \varepsilon_1$, where $\varepsilon_1, \varepsilon_2 > 0$, so that the unlabelled data contribute to the gradient.

2. The perturbation also tends to prolong the training of the discriminator. For a perfect GAN, with the training going on, the generator captures the data distribution. The discriminator can not distinguish the two distributions, which means $p_{\boldsymbol{x}_f} = p_{\boldsymbol{x}_{\text{ul}}}$, and the discriminator probability becomes $D(\boldsymbol{x}) \approx \frac{1}{2}$ [13]. When the discriminator is polluted by information leaking, $p(y = i|\boldsymbol{x}_{\text{ul}})$ potentially decreases, and $p(y = K + 1|\boldsymbol{x}_f)$ increases: $p(y = i|\boldsymbol{x}_{\text{ul}}, y \leq K) \approx \frac{1}{2} - \varepsilon_1$ and $p(y = K + 1|\boldsymbol{x}_f) \approx \frac{1}{2} + \varepsilon_2$, where $\varepsilon_1, \varepsilon_2 > 0$ decrease gradually with the training going on. Thus the discriminator gets more opportunities to learn better feature representations.

### 3.4. Focal consistency

In retina vessel images, it is challenging to deal with micro-vessels with delicate structures as they may take up only a few pixels in width and the classification can be easily disrupted by noise. There is a severe class imbalance problem during training. Inspired by [24], we propose a focal consistency loss function which down-scales the weights of easy examples, and thus guides the training to focus on the hard negatives.

The focal loss function [24] tries to mitigate the effect of the cross entropy loss being overwhelmed by large class imbalance. Let $p_t$ denote the probability for class $t$. The relevant cross entropy $-\log p_t$ can be scaled by a weighting factor $\alpha_t$ (which can be set as the inverse of class frequency or as a hyperparameter). Adding a modulating factor $(1 - p_t)^\rho$, with $\rho > 0$ being a tunable focusing parameter, we have the balanced focal loss function defined as

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\rho \log p_t. \quad (12)$$

Based on this, we propose a novel *focal consistency loss function* as follows by considering the prediction difference between the student and teacher networks:

$$\mathcal{L}_{\text{cons-FL}} = -\alpha_t(|p_t(\boldsymbol{x}_{\text{ul}}) - p_t(\boldsymbol{x}'_{\text{ul}})|)^\rho[\log p_t(\boldsymbol{x}_{\text{ul}}) - \log p_t(\boldsymbol{x}'_{\text{ul}})] \quad (13)$$

where $|p_t(\boldsymbol{x}_{\text{ul}}) - p_t(\boldsymbol{x}'_{\text{ul}})|$ is the perturbation factor. $\log p_t(\boldsymbol{x}_{\text{ul}}) - \log p_t(\boldsymbol{x}'_{\text{ul}})$ is the prediction distance between the teacher and the student networks where the input signal and weights happened to perturbation. The distance of the

two predictions decides the perturbation factor. It modulates the training range according to the consistency loss. When an example is aligned not well, the weight is improved due to the larger difference. When the two predictions are similar, it downs the loss weight. As the same as the reference [24], the parameter adjusts the rate at which aligned-well examples are down-weighted.

The consistency loss can extend the training range when the two prediction results have a more massive shift. It overcomes the distribution mismatch, possibly caused by the different view angles, lighting conditions, or acquisition devices. The dynamically scaled predict discrepancy boosts the generalization of the discriminator to adapt the transfer learning tasks.

### 3.5. Learning in LeakGAN

Having illustrated our framework with all loss functions defined, we give the overall learning objective as follows:

$$\mathcal{L}^* = \min_G \max_D (\lambda_1 \mathcal{L}_{\text{sup}} + \lambda_2 \mathcal{L}_{\text{unsup}} + \lambda_3 \mathcal{L}_{\text{cons-FL}}) \quad (14)$$

where $\lambda_1, \lambda_2, \lambda_3$ are the trade-off hyper-parameters to control the regularization of the model weights. Our model is trained like a standard GAN. The discriminator (also a classifier) is optimized by the joint objective. According to the analysis in section 3.4, we also apply the focal loss to $\mathcal{L}_{\text{sup}}$ to tackle the imbalance problem during the training. Meanwhile, for the optimization of the generator, we observe that our model can be trained by the standard GAN's protocol empirically:

$$\mathcal{L}_{\text{adv-}} = \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})} \log(1 - p(D(G(\boldsymbol{z})))). \quad (15)$$

During the training, we optimize the models iteratively. Besides, the consistency of the mean-teacher regularizes the discriminator as its student network.

## 4. Experiments

### 4.1. Datasets

The proposed method is evaluated on three public datasets, DRIVE [40], STARE [17] and CHASE_DB1 [10]. These datasets provide corresponding vessel segmentation manually labelled by domain experts, which can be seen as the ground truth segmentation. The details of the datasets are as follows.
**DRIVE:** There are 40 color fundus images with a resolution of $565 \times 584$ pixels. These images were captured by a Canon CR5 non-mydriatic camera with a 45-degree field of view (FOV). The bit-depth is 8-bit.
**STARE:** It consists of 20 color fundus images, with half of them containing signs of various pathologies. These images were captured with the resolution of $700 \times 605$ pixels by a

Table 1: Performance evaluation for DRIVE, STARE and CHASE_DB1. Semantic segmentation accuracy (Acc), Sensitivity (Se), and Specificity (Sp) are considered for the evaluations. Best performance highlighted in bold.

| Dataset | DRIVE | | | STARE | | | CHASE_DB1 | | |
|---|---|---|---|---|---|---|---|---|---|
| Methods | Acc(%) | Sp(%) | Se(%) | Acc(%) | Sp(%) | Se(%) | Acc(%) | Sp(%) | Se(%) |
| MS-NFN [42] | 95.67 | 98.19 | 78.44 | - | - | - | 96.37 | **98.47** | 75.38 |
| DeepVessel [11] | 95.23 | - | 76.03 | 95.85 | - | 74.12 | 94.89 | - | 71.30 |
| SegmentLoss [44] | 95.42 | 98.18 | 76.53 | 96.12 | 98.46 | 75.81 | 96.10 | 98.09 | 76.33 |
| IterNet [22] | 95.73 | **98.38** | 77.35 | **97.01** | **98.86** | 77.15 | 96.55 | 98.23 | 79.70 |
| LeakGAN 3 | 93.31 | 88.52 | 94.31 | 94.83 | 95.19 | 84.58 | 94.20 | 80.78 | 94.25 |
| LeakGAN 5 | 94.69 | 83.14 | 97.00 | 94.89 | 88.21 | **96.32** | 96.65 | 92.03 | 92.25 |
| LeakGAN 8 | **95.74** | 86.72 | **97.50** | 95.65 | 91.86 | 91.02 | **96.83** | 92.21 | **94.72** |

canon TopCon TRV-50 fundus camera with 35-degree FOV, and the bit-depth is 8-bit.

**CHASE_DB1:** There are 28 color fundus images from patients who participated in the child heart and health study in England. These images were captured with the resolution of $999 \times 960$ pixels by a Nidek NM 200D fundus camera with 30-degree FOV, and the bit-depth is 8-bit.

### 4.2. Quantitative Results

We evaluate the performance of our model on the three datasets and compare our model with state-of-the-art methods such as DeepVessel [11], SegmentLoss [44], MS-NFN [42] and IterNet [22]. Followed the protocol in [37], we use segmentation accuracy (Acc), specificity (Sp), sensitivity (Se) as quantitative results to measure the segmentation performance. They are defined as follows.

$$Se = \frac{TP}{TP + FN},$$

$$Sp = \frac{TN}{TN + FP},$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP},$$

where TP, true positives, are the vessel pixels predicted correctly. FP (false positive), TN (true negative), and FN (false negative) are similarly defined.

During the evaluations, the trained patches are cropped randomly. To evaluate the semi-supervised learning performance using different amounts of labelled data, we set the number of labelled training retina image to be 3, 5, and 8, respectively. The rest of the training images are treated as unlabelled samples. The evaluation result is shown in Table 1.

As seen, when using 3, 5 labelled images only ("LeakGAN 3" and "LeakGAN 5"), the performance of LeakGAN is already close to the comparison methods. When 8 labelled images are used ("LeakGAN 8"), its performance surpasses that of the counterparts with full supervised strategy on most of the criteria. Furthermore, our model has a balanced performance for the Se and Sp, while the results could benefit

from a higher Se value. The results show performance boost for all the three different tasks. It demonstrates that the unlabelled images can be effectively employed to train good classification boundaries, whereas more labelled images are beneficial for improved performance.

### 4.3. Qualitative Results

We proceed with the visual assessment of the semantic segmentation results. First, for DRIVE, whole-image predictions are shown in Fig. 3. Next, for more detailed visualization, patch predictions are also shown in Fig. 4. The results of the STARE and CHASE_DB1 are shown in Supplementary Fig. 6 and Fig. 7.

From the figures, we can see that our model predicts the retina fundus images well under the semi-supervised settings. They keep the preliminary information such as the retina's shape and the vessels' directions, though there are some differences. The differences in blue show the false positive pixels and yellow for false-negative ones. The blue pixels are mainly because of the extra vessel prediction alongside the real vessel, and the yellow pixels are due to a weak prediction value compared with the ground truth.

From the detailed patches prediction in Fig. 4, we can see that the thin vessels can be accurately detected from these three retina datasets. For example, in the third column in Fig. 4, the thin structure of the vessel is presented well by the prediction.

### 4.4. Results of Cross-Domain Semantic Segmentation

To evaluate our framework's generalization capacity, we conduct cross-domain semantic segmentation experiments following Algorithm 2 (see Supplementary). The proposed model is trained on one dataset and tested on another dataset. All the evaluations are under 8-labelled images setting. The results are shown in Table 2. In general, our model can generalize well to other fundus images captured with different cameras. The individual cross-dataset results are a little lower than the performance for within-dataset training and testing. Also, the performance from STARE or CHASE_DB1 to DRIVE is better than the opposite transfer
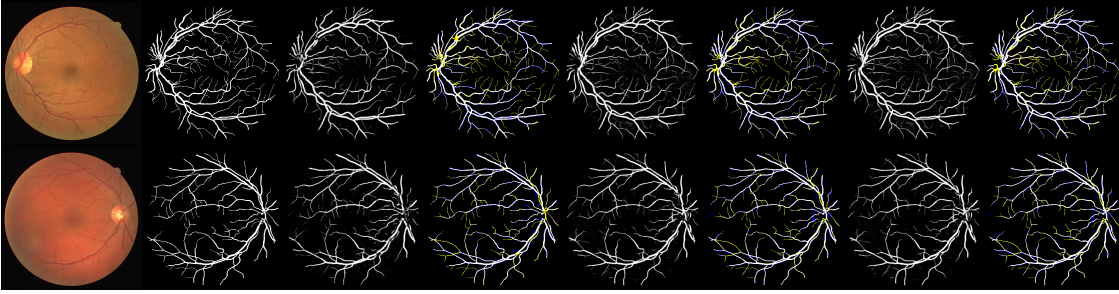
Figure 3: Visualization of the semantic segmentation on testing images for DRIVE. The results are shown when 3, 5 and 8 labelled training images are used. For the differences, the blue pixels being false positive, and yellow being false negative.
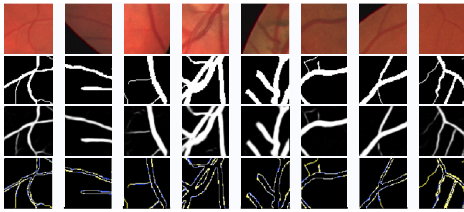


Figure 4: Patched segmentation results for the scenario of 8 labelled training image using DRIVE. The testing patches and their ground truth are in first and second row respectively. The third row is for the predictions. The differences are shown in the last row, blue pixels being false positive, and yellow being false negative.
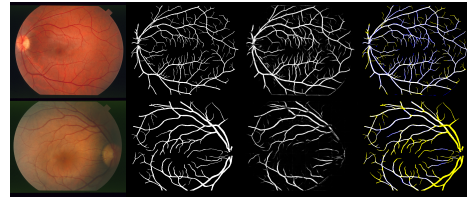


Figure 5: Visualization of the cross-domain prediction results for DRIVE to STRAE. For the difference, the blue pixels indicating false positive, and yellow indicating false negative.

performance. It might be due to the higher resolutions of STARE and CHASE_DB1.

Fig. 5 is the visualization for the evaluations of cross-domain segmentations from DRIVE to STARE, i.e., DRIVE is the training dataset, and STARE is the test dataset. The STARE vessels' entire structure still keeps clearly though some predictions are weak compared with the ground truth (yellow pixels in the figure). We also evaluate our model on the STARE to DRIVE and healthy DRIVE to pathological DRIVE. More detailed results for these different scenarios are shown in Supplementary.

### 4.5. Model Analysis

**Ablation studies:** To better understand the contribution made by the modules of the proposed LeakGAN model, we carry out an ablation study for each module. All the experiments are designed under a semi-supervised setting.

- U-Net: only labelled images are used for the U-Net training and testing;

- U-Net + GAN: GAN-based semi-supervised learning strategy is involved;

- U-Net + GAN + MT: Mean-teacher is further added;

- U-Net + GAN + MT + LM: Finally, the full model with information leaking involved.

We also experiment with using fewer labelled and unlabelled samples compared with the setting in Table 1. From Table 3, we can see that the U-Net can give good performance with few labelled retina images. However, it is prone to overfit, especially when only two samples are provided. The methods augmented by GAN and further by other modules perform increasingly better. With fixed amount of unlabelled images, more labelled images achieve higher performance. On the other hand, using more unlabelled data also gives higher accuracy for the three datasets. The full framework follows the same trends, but its overall performance across all datasets is the best among all models in comparison.

**Different information leaking settings:** According to our model's structure, there are four outputs from the generator that can be concatenated to the upsampling path of discriminator (from high-level features to low-level features) in different ways. The first injection with a shape of $8 \times 8$, then the size is doubled for the following ones up to $64 \times 64$. We evaluate three different leaking outputs (1st, 2nd, and 3rd intermediate outputs of the generator). Table 4 shows the results with different concatenated style. During the experiments, the $\alpha$ parameter is set to 1. From the table, we can see

Table 2: Performance of cross-domain evaluation. Semantic segmentation accuracy (Acc), Sensitivity (Se), and Specificity (Sp) are considered for the evaluations

| Training | DRIVE | | | | | | STARE | | | | | | CHASE_DB1 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Testing | STARE | | | CHASE_DB1 | | | DRIVE | | | CHASE_DB1 | | | DRIVE | | | STARE | | |
| Methods | Acc(%) | Sp(%) | Se(%) | Acc(%) | Sp(%) | Se(%) | Acc(%) | Sp(%) | Se(%) | Acc(%) | Sp(%) | Se(%) | Acc(%) | Sp(%) | Se(%) | Acc(%) | Sp(%) | Se(%) |
| EnsembleModel [10] | 94.95 | 97.70 | 70.10 | - | - | - | 94.56 | 97.92 | 72.42 | 94.15 | 96.65 | 71.03 | - | - | - | - | - | - |
| Cross-Mod [23] | 95.45 | **98.28** | 70.27 | **94.29** | **97.91** | 71.18 | 94.86 | **98.10** | 72.73 | 94.17 | **97.68** | 72.40 | 94.84 | **98.11** | 73.07 | **95.36** | **98.31** | 69.44 |
| BTS-DSN [14] | 95.48 | 98.16 | 71.88 | 94.41 | 97.15 | 69.80 | **95.02** | 97.84 | 74.46 | 94.11 | 97.10 | 67.26 | 93.77 | 96.99 | 69.60 | 95.01 | 98.08 | 67.99 |
| LeakGAN 8 | **95.53** | 93.30 | **85.69** | 94.25 | 96.43 | **72.65** | 94.35 | 83.21 | **90.25** | **94.28** | 88.25 | **86.23** | **94.95** | 89.16 | **86.85** | 94.68 | 80.30 | **83.23** |

Table 3: Ablation results under the different labelled and unlabelled settings; L: Labelled, UL: Unlabelled.

| Methods | Dataset | (2L, 5UL) | | | (5L, 5UL) | | | (2L, 6UL) | | | (2L, 8UL) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc(%) | Sp(%) | Se(%) | Acc(%) | Sp(%) | Se(%) | Acc(%) | Sp(%) | Se(%) | Acc(%) | Sp(%) | Se(%) |
| U-Net | DRIVE | 92.86 | 95.66 | 80.59 | 93.85 | 89.31 | 91.01 | 92.86 | 95.66 | 80.95 | 92.86 | 95.66 | 80.59 |
| | STARE | 93.16 | 83.67 | 94.60 | 94.87 | 89.74 | 90.03 | 93.16 | 83.67 | 94.60 | 93.16 | 83.67 | 94.60 |
| | CHASE_DB1 | 94.83 | 95.19 | 83.21 | 95.63 | 92.65 | 88.72 | 94.83 | 95.19 | 83.21 | 94.83 | 95.19 | 83.21 |
| U-Net+GAN | DRIVE | 94.26 | 87.65 | 96.50 | 94.36 | 89.21 | 93.20 | 94.61 | 75.83 | 94.86 | 93.54 | 80.45 | 93.13 |
| | STARE | 93.98 | 88.51 | 86.17 | 93.86 | 88.23 | 91.58 | 92.32 | 90.97 | 80.93 | 93.71 | 90.06 | 89.63 |
| | CHASE_DB1 | 94.88 | 90.62 | 89.09 | 94.66 | 95.21 | 84.65 | 94.83 | 90.74 | 89.18 | 94.96 | 88.61 | 89.96 |
| U-Net+GAN +MT | DRIVE | 94.05 | 87.63 | 91.58 | 94.48 | 80.75 | 95.76 | 94.86 | 87.53 | 92.67 | 93.54 | 80.28 | 93.55 |
| | STARE | 94.08 | 89.21 | 86.06 | 93.78 | 83.94 | 90.74 | 93.57 | 88.64 | 85.85 | 93.86 | 90.04 | 83.08 |
| | CHASE_DB1 | 94.10 | 83.68 | 94.51 | 93.84 | 82.23 | 95.57 | 94.30 | 91.50 | 86.21 | 94.38 | 85.51 | 92.08 |
| U-Net+GAN +MT+LM | DRIVE | 94.56 | 88.52 | 97.85 | 94.61 | 83.14 | 96.34 | 93.15 | 84.35 | 95.98 | 94.55 | 86.96 | 94.62 |
| | STARE | 94.32 | 85.17 | 96.31 | 94.89 | 88.21 | 92.32 | 93.73 | 89.74 | 90.23 | 94.05 | 89.31 | 95.12 |
| | CHASE_DB1 | 94.85 | 87.16 | 90.49 | 96.67 | 92.47 | 95.68 | 94.87 | 86.19 | 93.36 | 94.56 | 86.17 | 92.27 |

Table 4: Performance of different leaking module setting

| Dataset | Type 1 | Acc(%) | Sp(%) | Se(%) | Type2 | Acc(%) | Sp(%) | Se(%) | Type3 | Acc(%) | Sp(%) | Se(%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DRIVE | 1st | 95.72 | 86.72 | 97.50 | 1st+2nd | 94.63 | 80.32 | 97.72 | 1st+2nd+3rd | 93.58 | 81.40 | 98.18 |
| | 2nd | 93.83 | 83.62 | 96.91 | 1st+3rd | 94.39 | 86.53 | 95.56 | | | | |
| | 3rd | 94.30 | 85.48 | 97.12 | 2nd+3rd | 94.69 | 85.79 | 97.61 | | | | |

that the performance for different leaking settings is similar, showing no strong sensitivity to the leaking schemes.

We have also evaluated the effect of using different $\alpha$ values for the leaking module. Take the DRIVE experiment as an example as shown in Table 5; when $\alpha = 1$, the accuracy is the highest. Based on these evaluations suggest that we can choose to leak the first convolutional outputs of the generator and set the $\alpha$ to 1 for the leaking module.

Table 5: Performance of different $\alpha$ values for the leaking module

| $\alpha$ value | 0.1 | 0.5 | 1.0 | 1.1 | 1.5 |
|---|---|---|---|---|---|
| Acc(%) | 94.13 | 94.20 | 95.72 | 94.61 | 94.67 |
| Sp(%) | 80.54 | 82.48 | 86.72 | 80.13 | 81.47 |
| Se(%) | 98.10 | 98.45 | 97.50 | 97.56 | 98.30 |

**Analysis of the focal consistency loss:** To evaluate our focal consistency's performance, we first investigate the different $\rho$ and $\alpha_t$. We also compare it with traditional consistency based on MSE in DRIVE. During the experiments, we follow the protocol in [24]. The results are shown in Table 6. We can see that the best performance occurs with $\alpha_t = 2.0$ and $\rho = 0.25$.

Table 6: Evaluations of the $\alpha_t$ and $\rho$ of focal loss for DRIVE.

| Focal Loss | | Acc(%) | Sp(%) | Se(%) |
|---|---|---|---|---|
| $\alpha_t$ | $\rho$ | | | |
| 1.0 | 0.25 | 92.10 | 83.56 | 93.72 |
| 2.0 | 0.25 | 95.72 | 86.72 | 97.50 |
| 5.0 | 0.25 | 94.69 | 83.84 | 97.53 |
| MSE | | 94.47 | 85.63 | 96.15 |

## 5. Conclusion

In this paper, we have proposed a novel semi-supervised semantic segmentation method for retina vessel data. To address the issue of lacking labelled data for training a complex semantic segmentation model, a GAN-based framework is employed that integrates information leaking and mean-teacher mechanisms. Information leaking from the generator to the discriminator pollute the generated examples, which helps to improve the generalization performance. The mean-teacher regularization is used to cope with variations in retina images due to changes in imaging conditions. Furthermore, to address the class imbalance problem, a novel focal consistency loss is proposed for the mean-teacher regularization.

The generalization ability of our proposed model has been demonstrated by extensive empirical validations using three widely used retina image datasets. We will apply our model to other kinds of medical images in our future work.

# References

[1] Mart Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467 [cs.DC]*, 2016. 11

[2] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[3] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 810–818, 2019. 1, 2

[4] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson. Robust vessel segmentation in fundus images. *International Journal of Biomedical Imaging (IJBI)*, 2013:1–11, 2013. 1

[5] Qi Cai, Yingwei Pan, Chong-Wah Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11457–11466, 2019. 2, 4

[6] Liang Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 833–851, 2018. 1

[7] Xu Chen, Bryan M Williams, Srinivasa R Vallabhaneni, Gabriela Czanner, Rachel Williams, and Yalin Zheng. Learning active contour models for medical image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11624–11632, 2019. 1

[8] Zihang Dai, Zhilin Yang, Fan Yang, William W. Cohen, and Ruslan Salakhutdinov. Good semi-supervised learning that requires a bad GAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6511–6521, 2017. 2, 4

[9] M. M. Fraz, S. A. Barman, P. Remagnino, A. Hoppe, A. Basit, B. Uyyanonvara, A. R. Rudnicka, and C. G. Owen. An approach to localize the retinal blood vessels using bit planes and centerline detection. *Computer Methods and Programs in Biomedicine*, 108(2):600–616, 2012. 1

[10] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R. Rudnicka, Christopher G. Owen, and Sarah A. Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Bio-Medical Engineering*, 59:2538–2548, 2012. 2, 5, 8

[11] Huazhu Fu, Yanwu Xu, Stephen Lin, Damon Wing Kee Wong, and Jiang Liu. Deepvessel: Retinal vessel segmentation via deep learning and conditional random field. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 132–139, 2016. 2, 6

[12] Zhe Gan, Liqun Chen, Weiyao Wang, Yunchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin. Triangle generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5247–5256, 2017. 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems (NeurIPS)*, pages 2672–2680, 2014. 5

[14] Song Guo, Kai Wang, Hong Kang, Yujun Zhang, Yingqi Gao, and Tao Li. BTS-DSN: Deeply supervised neural network with short connections for retinal vessel segmentation. *International Journal of Medical Informatics*, 126:105–113, 2019. 2, 8

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[16] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1495–1503, 2015. 2

[17] Adam Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging (TMI)*, 19(3):203–210, 2000. 2, 5

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 2

[19] Sharif Amit Kamran, Khondker Fariha Hossain, Alireza Tavakkoli, Stewart Lee Zuckerbrod, Kenton M. Sanders, and Salah A. Baker. RV-GAN: segmenting retinal vascular structure in fundus photographs using a novel multi-scale generative adversarial network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 34–44, 2021. 2

[20] Abhishek Kumar, Prasanna Sattigeri, and P. Thomas Fletcher. Semi-supervised learning with GANs: Manifold invariance with improved inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5535–5545, 2017. 2, 4

[21] Chen-Yu Lee, Saining Xie, Patrick W Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In Guy Lebanon and S V N Vishwanathan, editors, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *PMLR*, pages 562–570, 2015. 2

[22] Liangzhi Li, Manisha Verma, Yuta Nakashima, Hajime Nagahara, and Ryo Kawasaki. Iternet: Retinal image segmentation utilizing structural redundancy in vessel networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3645–3654, 2020. 2, 6

[23] Qiaoliang Li, Bowei Feng, LinPei Xie, Ping Liang, Huisheng Zhang, and Tianfu Wang. A cross-modality learning approach for vessel segmentation in retinal images. *IEEE Transactions on Medical Imaging (TMI)*, 35(1):109–118, 2015. 2, 8

[24] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 5, 8

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 2

[26] Michael Majurski, Petru Manescu, Sarala Padi, Nicholas Schaub, Nathan Hotaling, Carl Simon, and Peter Bajcsy. Cell image segmentation using generative adversarial networks, transfer learning, and augmentations. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 1114–1122, 2019. 1

[27] Sara Moccia, Elena De Momi, Sara El Hadji, and Leonardo S Mattos. Blood vessel segmentation algorithms — review of methods, datasets and vvaluation metrics. *Computer Methods and Programs in Biomedicine*, 158:71–91, 2018. 1, 4

[28] Arnab Kumar Mondal, Jose Dolz, and Christian Desrosiers. Few-shot 3D multi-modal medical image segmentation using generative adversarial learning. *arXiv:1810.12241 [cs.CV]*, 2018. 11

[29] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10):1345–1359, 2010. 2

[30] Kyeong-Beom Park, Sung Ho Choi, and Jae Yeol Lee. M-GAN: retinal blood vessel segmentation by balancing losses through stacked deep fully convolutional networks. *IEEE Access*, 8:146308–146322, 2020. 2

[31] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016. 11

[32] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3546–3554, 2015. 2

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 1, 2

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2234–2242, 2016. 1, 2, 3

[35] Firas Shama, Roey Mechrez, Alon Shoshan, and Lihi Zelnik-Manor. Adversarial feedback loop. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3205–3214, 2019. 2

[36] Seung Yeon Shin, Soochahn Lee, Il Dong Yun, and Kyoung Mu Lee. Deep vessel segmentation by learning graphical connectivity. *Medical Image Analysis (MIA)*, 58:101556, 2019. 1, 2

[37] Toufique Ahmed Soomro, Ahmed J. Afifi, Lihong Zheng, Shafiullah Soomro, Junbin Gao, Olaf Hellwich, and Manoranjan Paul. Deep learning models for retinal blood vessels segmentation: A review. *IEEE Access*, 7:71696–71717, 2019. 2, 6

[38] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5688–5696, 2017. 2

[39] Rupesh Kumar Srivastava, Klaus Greff, and J´urgen Schmidhuber. Training very deep networks. In *Advances In Neural Information Processing Systems (NeurIPS)*, pages 2377–2385, 2015. 2

[40] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging (TMI)*, 23(4):501–509, 2004. 2, 5

[41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1196–1205, 2017. 1, 2, 4

[42] Yicheng Wu, Yong Xia, Yang Song, Yanning Zhang, and Weidong Cai. Multiscale network followed network model for retinal vessel segmentation. In Alejandro F Frangi, Julia A Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 119–126, 2018. 6

[43] Rui Xu, Tiantian Liu, Xinchen Ye, Lin Lin, and Yen-Wei Chen. Boosting connectivity in retinal vessel segmentation via a recursive semantics-guided network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 786–795, 2020. 2

[44] Zengqiang Yan, Xin Yang, and Kwang Ting Cheng. Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation. *IEEE Transactions on Biomedical Engineering (TBE)*, 65(9):1912–1923, 2018. 1, 2, 6

[45] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2074–2083, 2019. 1, 2

[46] Yuyin Zhou, Yan Wang, Peng Tang, Song Bai, Wei Shen, Elliot K Fishman, and Alan Yuille. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 121–140, 2019. 1, 2