Accuracy Improvement of Semi-Supervised Segmentation Using Supervised ClassMix and Sup-Unsup Feature Discriminator

Takahiro Mano[a], Reiji Saito[b] and Kazuhiro Hotta[c]
Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan

Keywords: Semi-Supervised Learning, Segmentation, SupMix, ClassMix.

Abstract: In semantic segmentation, the creation of pixel-level labels for training data incurs significant costs. To address this problem, semi-supervised learning, which utilizes a small number of labeled images alongside unlabeled images to enhance the performance, has gained attention. A conventional semi-supervised learning method, ClassMix, pastes class labels predicted from unlabeled images onto other images. However, since ClassMix performs operations using pseudo-labels obtained from unlabeled images, there is a risk of handling inaccurate labels. Additionally, there is a gap in data quality between labeled and unlabeled images, which can impact the feature maps. This study addresses these two issues. First, we propose a method where class labels from labeled images, along with the corresponding image regions, are pasted onto unlabeled images and their pseudo-labeled images. Second, we introduce a method that trains the model to make predictions on unlabeled images more similar to those on labeled images. Experiments on the Chase and COVID-19 datasets demonstrated an average improvement of 2.07% in mIoU compared to conventional semi-supervised learning methods.

1 INTRODUCTION

In recent years, with advancements in image recognition technology, various models have been proposed, such as the fully convolutional network FCN (Long et al., 2015), encoder-decoder structures like SegNet (Badrinarayanan et al., 2017) and U-Net (Ronneberger et al., 2015), and the more advanced Deeplabv3+ (Chen et al., 2018). However, a large amount of labeled data is generally required when performing image recognition using deep learning. Among these tasks, semantic segmentation is particularly demanding as it requires pixel-level labeling, making the preparation of a large dataset of labeled images costly.

In recent years, when using a large amount of labeled images, it has become possible to achieve high accuracy. However, when training with only a small number of labeled images, the accuracy significantly decreases. In real-world applications, it is desirable to reduce high costs by achieving high accuracy with only a limited number of labeled images. Against the background, a learning method called semi-supervised learning, which utilizes a small amount of labeled images alongside unlabeled images for model training, has garnered attention.

In semi-supervised segmentation, a technique called pseudo-labeling (Lee et al., 2013) is the dominant approach. Semi-supervised segmentation heavily relies on the quality of these pseudo-labels. In medical imaging, which is the focus of this paper, there is often a significant class imbalance, making it difficult to predict rarely occurring classes, which in turn degrades the quality of pseudo-labels. Furthermore, since the model learns by treating the predicted pseudo-labels as ground truth, incorrect learning can lead to decreased accuracy. Therefore, when learning rare classes, it is crucial to ensure proper learning in the limited opportunities available.

A conventional semi-supervised segmentation method is ClassMix (Olsson et al., 2021). ClassMix involves cutting out the region of a randomly selected half of the predicted classes (e.g., one class if there are two) from one image and pasting it onto another image. By considering the shape of the class during the cut-and-paste process, ClassMix helps the model learn semantic boundaries between classes more effectively. However, ClassMix has two major issues. The first issue is that it performs ClassMix using pre-

[a] https://orcid.org/0000-0003-2077-6079
[b] https://orcid.org/0009-0003-5197-8922
[c] https://orcid.org/0000-0002-5675-8713

Mano, T., Saito, R. and Hotta, K.
Accuracy Improvement of Semi-Supervised Segmentation Using Supervised ClassMix and Sup-Unsup Feature Discriminator.
DOI: 10.5220/0013222700003912
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 3: VISAPP*, pages 592-600
ISBN: 978-989-758-728-3; ISSN: 2184-4321

dictions on unlabeled images. When ClassMix uses the prediction results for unlabeled images, the accuracy of the mixed images becomes dependent on the model's prediction accuracy for unlabeled images. If the model makes incorrect predictions, the quality of the mixed images deteriorates. The second issue is that the regions of half the classes are selected randomly. Especially, since the background class has a large number of samples, it is likely to already have high accuracy. Therefore, learning by pasting classes that are already predicted with high accuracy does not provide much benefit. Additionally, rare classes are less likely to become pseudo-labels because their prediction confidence remains low until learning progresses. This approach is not effective in datasets with significant class imbalances. To address these two issues, we propose a method called Supervised ClassMix (SupMix).

SupMix mixes regions except for background class, which has a large number of samples from labeled images, into pseudo-labels from unlabeled images. By attaching labels from different labeled images to the pseudo-labels, the accuracy of the pseudo-labels is improved without relying on the model's prediction accuracy, thereby addressing the issue of low accuracy in the initial pseudo-labels. Furthermore, by pasting regions other than background class from labeled images onto different unlabeled images, class imbalance can be mitigated. This helps to address the second issue of class imbalance.

In the research on semi-supervised learning, the domain gap between labeled and unlabeled images is often not considered. However, in real-world scenarios, there is an abundance of unlabeled images. If this domain shift can be properly addressed, semi-supervised learning can integrate more knowledge from unlabeled images. Therefore, this study focuses on minimizing the domain gap between predictions from labeled and unlabeled images. Specifically, we use a Generative Adversarial Network (GAN) (Goodfellow et al., 2014) to train the model so that the feature maps obtained from labeled images and those from unlabeled images are indistinguishable. By reducing the domain gap between the features extracted from labeled and unlabeled images, the model can efficiently acquire information from the unlabeled data, ultimately leading to improve the accuracy.

We conducted experiments on the Chase (Guo et al., 2021; Fraz et al., 2012) and COVID-19 (QMENTA, 2020) datasets. Our goal was to improve the accuracy over UniMatch (Yang et al., 2023), a good precision method in semi-supervised segmentation. We compared the performance of UniMatch with our proposed method under conditions where only 1/4 and 1/8 of the total labeled images were used. Across all datasets, our proposed method achieved higher mIoU than the UniMatch. For the Chase dataset, when we use 1/8 of the total labeled images, the IoU for the class with a small number of samples "retinal vessel" improved by 3.30% compared to UniMatch. When 1/4 of the labeled images is used, the IoU for "retinal vessel" increased by 2.63%. In the COVID-19 dataset, the IoU for the the class with a small number of samples "ground-glass" improved by 10.7% when we use 1/8 of the labeled images, and by 4.76% compared to UniMatch when 1/4 of the labeled images is used.

The structure of this paper is as follows: In Section 2, we discuss related researches. Section 3 explains the details of the proposed method. In Section 4, we present experimental results and provide a discussion. Finally, Section 5 concludes the paper and outlines future challenges.

2 RELATED WORKS

2.1 Consistency Regularization

Consistency regularization frameworks (Jeong et al., 2019; Chen et al., 2021b; Zou et al., 2021) are based on the idea that the predictions of unlabeled images should remain invariant even after applying augmentations. A common technique in classification tasks is called augmentation anchoring. Consistency regularization involves training in such a way that the predictions of augmented samples are forced to be consistent with the predictions the original unaugmented images. Our method utilizes this augmentation anchoring technique. The model is trained to maintain consistency between pseudo-labeled images, which are predictions of unlabeled images with weak augmentations, and synthetic images created by pasting the class shape of labeled images onto the pseudo-labeled images (SupMix). By using labeled images, the accuracy of the labels improves, ultimately leading to better overall performance.

2.2 Augmentation Methods

The Cutout (Devries and Taylor, 2017) algorithm is a technique that masks a square region within an image. By hiding specific partial areas, the model is encouraged not to rely on any particular region, allowing for a better understanding of the overall meaning of the image. The Random Erasing (Zhong et al., 2020) algorithm, on the other hand, removes random rectangular areas. Unlike Cutout, it does not restrict itself to squares, and the size and location of the erased regions are determined randomly. The Mixup (Zhang et al., 2018) algorithm is a method that linearly mixes two images and their corresponding labels, enabling the model to learn intermediate representations and become more robust to diverse data. The CutMix (Yun et al., 2019) algorithm blends two different images by cutting out a random rectangular region from one image and pasting it onto another. CutMix also mixes the labels of both images based on the proportion of the rectangular area.

ClassMix (Olsson et al., 2021) is an augmentation technique where randomly selected classes predicted from one image are cut out and pasted onto another image. Unlike CutMix, which cuts out a random rectangular region, leading to differences in context between the cut-out image and the destination image, making learning more difficult, ClassMix considers the shape of the class when cutting and pasting. This allows the model to learn the semantic boundaries of each class more effectively. However, conventional ClassMix relies on predictions from unlabeled images, which introduces the problem of depending on the prediction accuracy of those unlabeled images. To address this issue, we propose to paste a small number of classes from labeled images instead of relying on predictions from unlabeled images, which helps maintain the quality of pseudo-labels.

2.3 Adversarial Methods

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are used in various tasks beyond image generation, such as semantic segmentation (Chen et al., 2021a; Tsuda and Hotta, 2019) and appearance inspection (Schlegl et al., 2017; Zenati et al., 2018; Akcay et al., 2019), and have achieved strong performance. In the field of semi-supervised semantic segmentation, several methods leveraging GANs have also been proposed.

The first adversarial approach (Souly et al., 2017) used in semi-supervised semantic segmentation involves the generator increasing the number of samples available for training, while the discriminator also acts as the segmentation network. The output of the discriminator classifies each pixel as belonging to a correct class or a fake class. This enables the segmentation network to improve the ability to distinguish between real (supervised and unsupervised samples) and generated samples. By treat supervised and unsupervised samples as the same class in discriminator, the method makes close the features of supervised and unsupervised samples indirectly.

However, none of the existing methods directly consider the domain gap between labeled and unlabeled images. Therefore, we aim to reduce the domain gap between the predictions of labeled and unlabeled images. To achieve this, we pass both labeled and unlabeled images through the model to obtain feature maps and then feed them into a discriminator, training it to distinguish between the two. The segmentation model is trained in such a way that it becomes difficult to differentiate whether the feature maps are from labeled or unlabeled images. This approach allows us to extract rich information from a large amount of unlabeled images, and we believe that it leads to improve the accuracy.

3 PROPOSED METHOD

We focus on semi-supervised segmentation, particularly with imbalanced medical datasets. We attempted two improvements. The first challenge is a modification of ClassMix (Olsson et al., 2021) described in Section 3.1. The second challenge is to address the domain gap between the feature maps of labeled and unlabeled images. This is explained in Section3.2.

3.1 Supervsed ClassMix(SupMix)

We focus on enhancing ClassMix. First, as a preliminary step for the subsequent improvements, we introduce ClassMix and clarify the issues. As shown on the left of Figure 1, ClassMix is a technique where half of the predicted classes from one image are randomly selected, and their corresponding regions are cut out and pasted onto another image. This method cuts and pastes regions while considering the shapes of the classes, allowing for more accurate learning of the semantic boundaries of each class. We will now introduce the ClassMix algorithm. First, we prepare two unlabeled images, $x_A \in \mathbb{R}^{3\times H\times W}$ and $x_B \in \mathbb{R}^{3\times H\times W}$, where H and W represent the height and width of the images. The two unlabeled images x_A and x_B are then fed into a model f (such as DeepLabv3+ (Chen et al., 2018) specialized for segmentation).

$$y_A = Argmax_c(f(x_A)) \tag{1}$$
$$y_B = Argmax_c(f(x_B)) \tag{2}$$

where $y_A \in \mathbb{R}^{H\times W}$ and $y_B \in \mathbb{R}^{H\times W}$ represent the predicted class labels obtained by passing the input images through the model. Additionally, we retrieve the number of classes $\hat{C}$ present in y_A and randomly select half of those classes. For an even number of classes (e.g., 2 classes), 1 class is selected. For an odd number of classes (e.g., 3 classes), half of the classes are

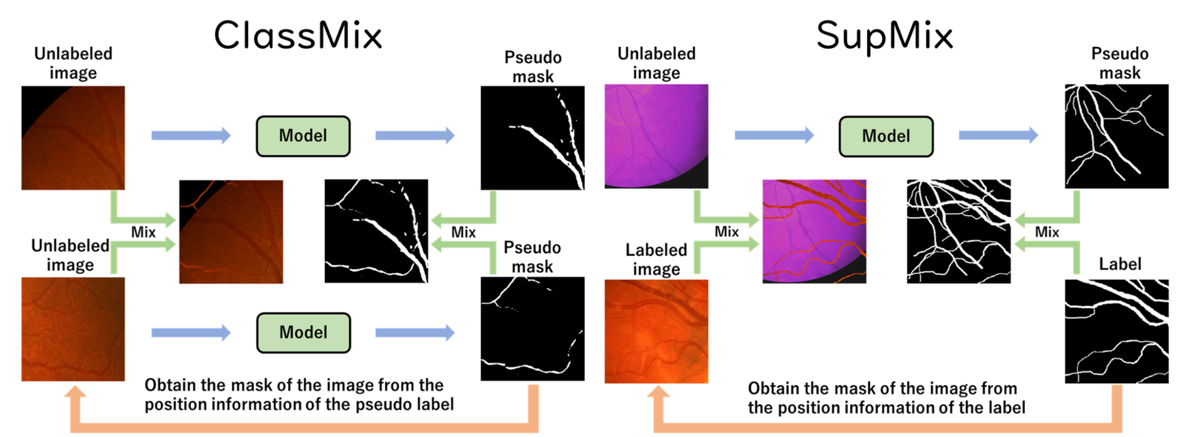
ClassMix
Unlabeled image
Model
Pseudo mask
Mix
Unlabeled image
Mix
Pseudo mask
Model
Obtain the mask of the image from the position information of the pseudo label
SupMix
Unlabeled image
Model
Pseudo mask
Mix
Labeled image
Mix
Label
Obtain the mask of the image from the position information of the label

Figure 1: Comparison between the conventional ClassMix and the proposed SupMix. ClassMix is mixing images between images and their segmentation predictions. In contrast, SupMix pastes regions from a few classes in labeled images onto the pseudo-labels of unlabeled images. The accuracy of the pseudo-labels mixed by SupMix is improved by using labeled images and pasting them onto unlabeled images with pseudo-labels. Since the ground-truth labels are independent of the prediction accuracy, only a few classes must be pasted onto the unlabeled images and pseudo-labels.

selected by discarding the decimal part (e.g., 1 class).

$$\hat{c} = \frac{\hat{C}}{2} \tag{3}$$

By using only the selected class $\hat{c}$ from the pseudo-label y_A of the unlabeled input image x_A, a binary mask $M_A \in \mathbb{R}^{H \times W}$ is generated.

$$\mathbf{M}_A = \begin{cases} 1 & \text{if } \mathbf{y}_A \in \hat{c} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Finally, by using the binary mask, we generate the synthesized input image $x_{mix} \in \mathbb{R}^{3 \times H \times W}$ and the synthesized pseudo-label $y_{mix} \in \mathbb{R}^{H \times W}$.

$$x_{mix} = M_A \odot x_A + (1 - M_A) \odot x_B \tag{5}$$

$$y_{mix} = M_A \odot y_A + (1 - M_A) \odot y_B \tag{6}$$

The synthesized input image x_{mix} and the synthesized pseudo-label y_{mix} represent the outputs of the ClassMix algorithm.

However, ClassMix has two issues. The first issue is that ClassMix is performed on both the unlabeled images and pseudo-labels. Since y_A and y_B equation 6 are pseudo-labels obtained from unlabeled input images, the accuracy of y_{mix} depends on the model's accuracy. As a result, if the model makes incorrect predictions, the accuracy of the mixed images will decrease (especially around the boundaries), making it difficult to improve the model's accuracy.

The second issue is that the classes for half the number of all classes are randomly selected and cut out for pasting. As shown in equation 3, randomly selecting half number of classes can result in pasting regions from classes with high accuracy into another image, which reduces the effectiveness of training even when high-accuracy classes are used. Furthermore, since pseudo-labels are used to select half of the classes, classes with other than background class tend to progress more slowly in training and are often not included in the pseudo-labels. For these reasons, ClassMix cannot adequately handle datasets with significant class imbalance. To address these two issues, we propose Supervised ClassMix (SupMix).

We present the overview of SupMix on the right side of Figure 1. Unlike ClassMix, which pastes pseudo-labels obtained from an unlabeled image onto another unlabeled image's pseudo-labels, SupMix mixes regions of specific classes obtained from a labeled image into the pseudo-labels of a weakly augmented unlabeled image. We define the labeled image and its label as x^l_A and y^l_A, respectively. These are subject to weak preprocessing such as cropping and horizontal flipping. The key difference from ClassMix is the use of labeled data. Additionally, an unlabeled image x^u_B is prepared, which is subjected to strong preprocessing (e.g., color jitter, blur, etc.). The unlabeled image x^u_B is passed through the model f to generate the pseudo-label y^u_B.

$$y^u_B = Argmax_c(f(x^u_B)) \tag{7}$$

Next, the number of existing classes C^l is obtained from the ground truth labels. SupMix allows selecting classes to paste from all classes in the ground truth label, whereas ClassMix can only paste classes predicted within the pseudo-label Background class is manually excluded, and we define this set as $C_{selected}$. The reason for manually excluding the background class compared to other classes is that it appears frequently during training due to its large number of samples, inducing its learning effectiveness. After excluding the background class, half of the remaining classes are randomly selected, which we define as $c^l_{selected}$. This serves as a solution to the second issue.

$$c^l_{selected} = \frac{C^l_{selected}}{2} \tag{8}$$

A binary mask $M_A \in \mathbb{R}^{H \times W}$ is generated from the pseudo-label y^l_A of the unlabeled input image x^l_A, containing only the selected class $c^l selected$.

$$\mathbf{M}^l_A = \begin{cases} 1 & \text{if } \mathbf{y}^l_A \in c^l_{selected} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$M^l_A \in \mathbb{R}^{H \times W}$ is a binary mask obtained from the ground truth labels. This ensures that the accuracy of the mask is always maintained when pasting it onto another image. As a result, the pasting process does not rely on the model's accuracy, allowing it to be applied directly to the unlabeled image. This serves as a solution to the first issue and is the primary advantage of using SupMix. Furthermore, with the modification of M^l_A, Equations 5 and 6 can be rewritten as follows.

$$x^l_{mix} = M^l_A \odot x^l_A + (1 - M^l_A) \odot x^u_B \tag{10}$$

$$y^l_{mix} = M^l_A \odot y^l_A + (1 - M^l_A) \odot y^l_B \tag{11}$$

The outputs x^l_{mix} and y^l_{mix} obtained from Equations 10 and 11 represent the final outputs of the SupMix algorithm. While there is a concern that pasting ground truth labels could lead to overfitting due to the repeated use of specific images or features, this is mitigated because the labeled images and their ground truth labels undergo preprocessing. Therefore, overfitting is considered less likely to occur.

3.2 Sup-Unsup Feature Discriminator

In conventional semi-supervised learning, the domain gap between labeled and unlabeled images is not considered. However, in real-world scenarios, there is an abundance of unlabeled images. If this domain shift can be handled appropriately, semi-supervised learning can incorporate more knowledge from unlabeled images. Therefore, this paper proposes the Sup-Unsup Feature Discriminator (SUFD) to reduce the domain gap between the predictions of labeled and unlabeled images.

Figure 2 provides the overview of the Sup-Unsup Feature Discriminator (SUFD). To reduce the domain gap between the predictions of labeled and unlabeled images, a discriminator commonly used in Generative Adversarial Networks (GAN), is employed. As shown in Figure 2, the discriminator is trained on the feature maps obtained from the model. The semi-supervised learning method, acting as a generator, is trained to produce feature maps from unlabeled images that the discriminator would mistake for features obtained from labeled images. In other words, this structure aims to make the feature maps derived from unlabeled images resemble features from labeled images, allowing the model to generate high-quality feature maps from unlabeled images similar to those from labeled ones. Additionally, the loss functions used to train the generator and discriminator are provided as

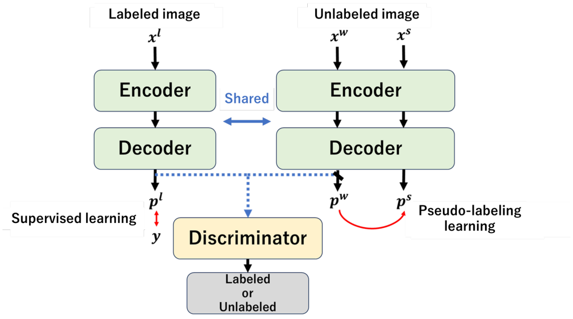

Figure 2: The Sup-Unsup Feature Discriminator involves a shared encoder-decoder architecture used for both supervised learning and pseudo-supervised learning. Note that x_l represents the labeled image, x_w represents the unlabeled image (weak augmentation), and x_s represents the unlabeled image (strong augmentation). p_l denotes the output feature map of the labeled image, p_w denotes the output feature map of the unlabeled image (weak augmentation), and p_s denotes the output feature map of the unlabeled image (strong augmentation). Additionally, y represents the ground truth. During training, a discriminator is employed to make it difficult to distinguish between the supervised features and the unsupervised features without fixing, which are obtained after applying weak augmentations.

$$\mathcal{L}_{SUFD} = B(o_u, 1) + \frac{1}{2}(B(o_u, 0) + B(o_l, 1)). \tag{12}$$

where $o_u \in \mathbb{R}^{c \times h \times w}$ is the unsupervised feature map output by the discriminator, while $o_l \in \mathbb{R}^{c \times h \times w}$ represents the supervised feature map. The first term is associated with learning the generator, while the second and third terms are related to training the discriminator. In the equation, B denotes the Binary Cross Entropy Loss. The discriminator performs a binary classification into two classes: 0 represents the "prediction from the unlabeled image", and 1 represents the prediction from the labeled image".

The generator is trained to make the discriminator misclassify unsupervised images as predictions from supervised images. This allows the feature maps obtained from unsupervised images to resemble those from supervised images. The discriminator is trained to output 0 when it identifies a prediction as being from an unsupervised image and 1 when it identifies it as being from a supervised image.

4 EXPERIMENTS

4.1 Datasets and Implementation Details

The Chase dataset consists of a total of 28 images. The dataset size is 999×960, and for this experiment, we used 23 images for training and 5 images for testing. The task is to predict four classes: background and retinal vessels. The pixel ratio of each class in the ground truth labels was measured, with the background class at 93.36% and the retail vessel class at 6.64%. From this, it is clear that there is a significant class imbalance.

The COVID-19 dataset consists of a total of 100 images, with 70 training images, 10 validation images, and 20 test images, all sized 256×256. The task is to predict four classes: background, ground-glass, consolidation, and pleural effusions. In this experiment, the validation images were not used; only the training and test images were utilized. When measuring the pixel ratio of each class in the ground truth labels, the background class is at 93.24%, the ground-glass class is at 2.14%, the consolidation class is at 4.50%, and the pleural effusions class is at 0.12%. From this, it is clear that the number of samples for classes other than the background class is significantly low.

We conduct experiments under conditions with limited supervision labels (semi-supervised learning). In this setting, we use 1/8 and 1/4 of the total training images as labeled images, while 7/8 and 3/4 are used as unlabeled images. We compare supervised learning, FixMatch (Sohn et al., 2020), UniMatch (Yang et al., 2023), and the proposed method(ours). The model used in this experiment is Deeplabv3+ (Chen et al., 2018) with a ResNet-101 backbone (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015).

The experiments were conducted using an Nvidia A6000. The batch size was set to 4, and the optimizer used was SGD with a momentum of 0.9 and a weight decay of $1 \times e^{-4}$. The initial learning rate for the scheduler was $4 \times e^{-3}$ for Chase and $1 \times e^{-3}$ for COVID-19, and it decayed after each iteration according to equation 13. The loss function used was Cross Entropy Loss. The model was trained for 1000 epochs on the Chase dataset and 500 epochs on the COVID-19 dataset.

Data augmentation in methods like FixMatch and UniMatch involves using weak and strong augmentations to learn from unlabeled images. The weak augmentations used include Random Crop and Random Horizontal Flip. The strong augmentations consist of these plus Random Color Jitter, Random Grayscale, Blur, and CutMix. The Random Crop is set to 320×320 for the Chase dataset and 256×256 for the COVID-19 dataset. The probability for Random horizontal Flip is set to 0.5. The probability for Random Color Jitter is 0.8, with brightness, contrast, and saturation all set to 0.5, and hue set to 0.25. The probability for Random Grayscale is 0.2, while the probabilities for Blur and CutMix are both set to 0.5. All of these settings are the same as in UniMatch. In FixMatch and UniMatch, it is possible to set a threshold for pseudo labels, which is set at 0.95. Additionally, UniMatch employs dropout with a probability of 0.5. The evaluation metric was conducted using Intersection over Union (IoU), and the evaluation was based on the average of the experimental results obtained by changing the initializations five times.

$$lr_{current} = lr_{init} \times \left(\frac{1 - \text{iteration}}{\text{epoch}} \right)^{0.9} \tag{13}$$

In the proposed method, instead of strong augmentation CutMix, Supervised ClassMix (SupMix) is used. SupMix allows for specifying class labels when performing pasting. For both Chase and COVID-19, apart from the class with a large number of samples (background), half of the other classes (1 class for both Chase and COVID-19) are selected uniformly.

4.2 Quantitative Evaluation

The top Table 1 shows the experiments on the Chase dataset. The proposed method (ours) outperformed UniMatch in both cases where the number of labeled images was 1/8 and 1/4. When we use 1/8 of the labeled images, our method achieved a 1.73% improvement in mIoU compared to UniMatch, with a notable 3.30% increase in accuracy for the retinal vessel class. When 1/4 of the labeled images is used, our method showed a 1.38% improvement in mIoU over UniMatch, with a 2.63% increase specifically for the retinal vessel class. The significant improvement in the retinal vessel class can be attributed to SupMix, which enhances learning opportunities for non-background areas by pasting pseudo-labels from different images. Additionally, the greater accuracy improvement with fewer labeled images is likely due to SUFD. SUFD is likely because the features of the unlabeled images

Table 1: Accuracy on Chase and COVID-19 datasets for Supervised, FixMatch, UniMatch, and ours. The top table shows results on Chase and the bottom table shows the results on COVID-19 dataset. Each row shows the IoU and standard deviation for each class, while each column compares the case where the labeled data covers 1/8 (N images) and 1/4 (N images) in the total dataset.

Chase	1/8 (2 images)				1/4 (5 images)			
	Supervised	FixMatch	UniMatch	ours	Supervised	FixMatch	UniMatch	ours
background	74.62$_{\pm 37.31}$	95.6$_{\pm 0.12}$	96.41$_{\pm 0.06}$	96.58$_{\pm 0.06}$	39.12$_{\pm 44.57}$	96.48$_{\pm 0.05}$	96.50$_{\pm 0.02}$	96.63$_{\pm 0.03}$
retinal vessel	1.86$_{\pm 2.54}$	38.54$_{\pm 2.86}$	55.61$_{\pm 0.81}$	58.91$_{\pm 0.31}$	3.78$_{\pm 3.09}$	57.62$_{\pm 0.60}$	57.78$_{\pm 0.35}$	60.41$_{\pm 0.20}$
mean IoU	38.24$_{\pm 17.51}$	67.07$_{\pm 1.49}$	76.01$_{\pm 0.43}$	77.74$_{\pm 0.18}$	21.45$_{\pm 23.83}$	77.05$_{\pm 0.32}$	77.14$_{\pm 0.18}$	78.52$_{\pm 0.10}$
COVID-19	1/8 (9 images)				1/4 (18 images)			
background	96.26$_{\pm 0.07}$	95.57$_{\pm 0.23}$	96.65$_{\pm 0.13}$	96.72$_{\pm 0.11}$	96.78$_{\pm 0.12}$	96.65$_{\pm 0.08}$	97.03$_{\pm 0.05}$	97.08$_{\pm 0.07}$
ground-glass	25.11$_{\pm 0.97}$	33.52$_{\pm 3.54}$	26.55$_{\pm 3.94}$	37.25$_{\pm 1.41}$	32.50$_{\pm 2.84}$	48.18$_{\pm 1.06}$	34.09$_{\pm 1.21}$	38.85$_{\pm 1.85}$
consolidation	39.74$_{\pm 2.32}$	0.0$_{\pm 0.0}$	49.12$_{\pm 1.65}$	50.86$_{\pm 1.35}$	45.12$_{\pm 3.88}$	0.0$_{\pm 0.0}$	47.57$_{\pm 1.45}$	50.93$_{\pm 1.14}$
pleural effusions	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$
mean IoU	40.28$_{\pm 0.78}$	32.27$_{\pm 0.87}$	43.08$_{\pm 0.75}$	46.21$_{\pm 0.60}$	43.60$_{\pm 1.51}$	36.21$_{\pm 0.28}$	44.67$_{\pm 0.63}$	46.71$_{\pm 0.68}$

effectively aligned with those of the labeled images, leading to the extraction of higher-quality features.

The bottom Table 1 shows the experiments on the COVID-19 dataset. The proposed method (ours) outperformed UniMatch in both cases where the number of labeled images was 1/8 and 1/4. When we use 1/8 of the labeled images, our method achieved a 3.13% improvement in mIoU compared to UniMatch. Notably, the ground-glass and consolidation classes other than the background class, achieved accuracy improvements of 10.7% and 1.74%. When 1/4 of the labeled images is used, our method showed a 2.04% improvement in mIoU over UniMatch. Notably, the ground-glass and consolidation classes other than the background class, achieved accuracy improvements of 4.74% and 3.36%. The reason for improving the accuracy is the same as the discussion we made for the Chase dataset.

4.3 Qualitative Evaluation

Figure 3 shows the segmentation results on the Chase and COVID-19 datasets. The areas highlighted in yellow indicate regions where accuracy has improved in comparison with UniMatch. In the Chase dataset, we see that the connectivity of retinal vessels has improved in the yellow areas for both 1/4 and 1/8 supervised images. This improvement is due to the SupMix method, where the supervised mask was pasted onto another image while maintaining the connectivity of the retinal vessels. Furthermore, the improvement in retinal vascular connectivity is larger for 1/8 than for 1/4 of all supervised images when we compare UniMatch with our method. This is because SUFD was able to gain more knowledge from the abundant unsupervised images. Compared to UniMatch, for 1/8 of all supervised images in the COVID-19 dataset, the area within the yellow frame is predicted well in the background. For 1/4 of all supervised images, we see that the proposed method is closer to the ground truth than any other method within the yellow frame. These results demonstrate that our technique (SupMix and SUFD) is superior compared to other methods.

4.4 Ablation Studies

We introduced the SupMix and SUFD methods, but we have not yet verified the individual effectiveness of each method. Therefore, we conducted experiments to evaluate each method individually. The experimental procedures were carried out in the same manner as in Section 4.1. The experimental results are shown in Table 2. The baseline is UniMatch. Comparisons were made with conventional augmentation methods: CutMix and ClassMix. SupMix and SUFD are our proposed methods, and the combination of these two is referred to as "ours" in the table. The experiments were conducted on the Chase and COVID-19 datasets, and results were obtained for cases where 1/8 and 1/4 of the fully supervised images were used. Additionally, only the accuracies for the classes with fewer samples (i.e., retinal vessels and ground-glass) are shown to verify if the methods address class imbalance effectively.

First, regarding SupMix, it achieved significant accuracy improvements across all experimental results compared to conventional methods such as CutMix and ClassMix. This improvement is likely due to the fact that the class shapes from the labeled images were directly pasted onto other images, effectively addressing class imbalance. For SUFD, a notable accuracy improvement was observed when using 1/8 of the fully supervised images compared to conventional UniMatch (CutMix), whereas less improvement was seen when we use 1/4 of the images. This is because a large amount of information could be obtained from the abundant unlabeled images, leading to the observed accuracy gains. Finally, by combining these

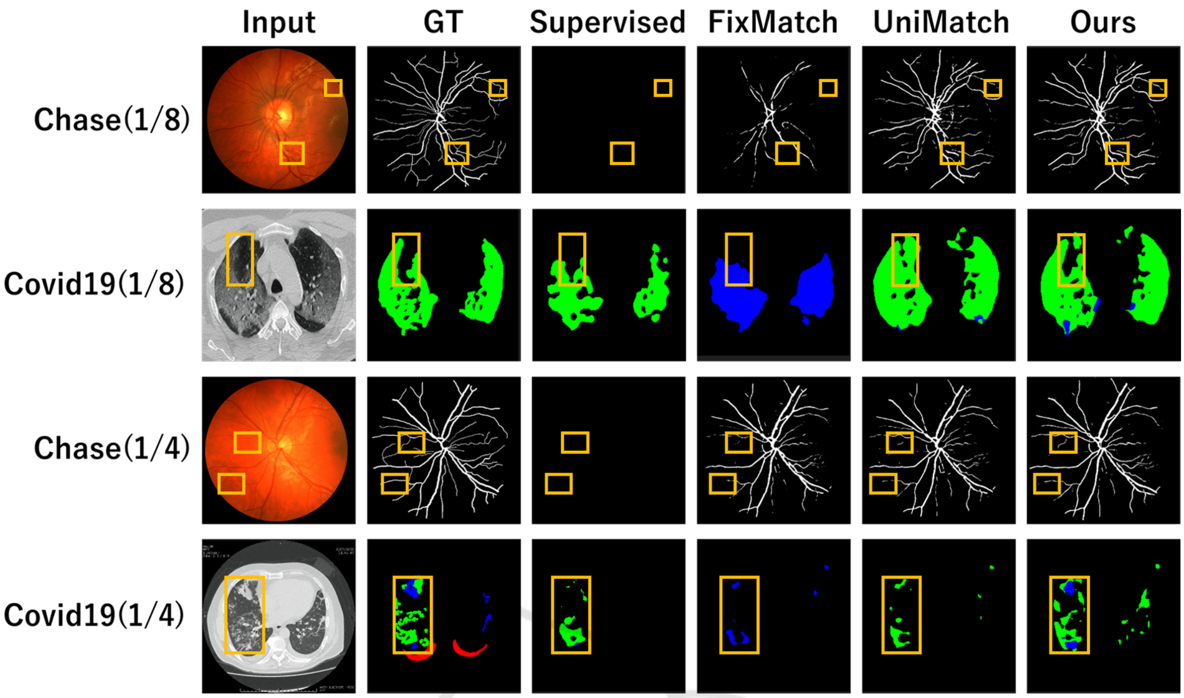

Figure 3: Segmentation results on Chase and COVID-19 datasets. Each row shows the dataset name and the ratio of supervised images used for training in all supervised images. Each column shows Input Images (Input), Ground Truth (GT), and the results of Supervised, FixMatch, UniMatch, and Ours. In the Chase dataset, the background class is visualized in black, and the retinal vessel class is in white. In the COVID-19 dataset, black represents the background class, blue indicates the ground-glass class, green shows the consolidation class, and red denotes the pleural effusions class.

Table 2: Verification of the individual effects of SupMix and SUFD. Each row indicates the dataset type and the classes with fewer samples (retinal vessel and ground-glass). In each column, the results display the accuracy when we train the network with 1/N images in all supervised images. The values in parentheses are the improvement over CutMix. The comparison methods include CutMix, ClassMix, SupMix, SUFD, and ours. CutMix corresponds to the standard UniMatch results, while "ours" refers to the combined method of SupMix and SUFD.

Methods	CutMix	ClassMix	SupMix	SUFD	ours	CutMix	ClassMix	SupMix	SUFD	ours
Chase	1/8 (2 images)					1/4 (5 images)				
retinal vessel	55.61(+0.00)	56.78(+1.17)	57.57(+1.96)	55.95(+0.34)	58.91(+3.30)	57.78(+0.00)	57.91(+0.13)	59.60(+1.82)	56.90(−0.88)	60.41(+2.63)
COVID-19	1/8 (9 images)					1/4 (18 images)				
ground-glass	26.55(+0.00)	27.95(+1.40)	32.82(+6.27)	33.36(+6.81)	37.25(+10.70)	34.09(+0.00)	34.99(+0.90)	38.10(+4.01)	34.17(+0.07)	38.85(+4.76)

methods, we achieved the best results overall.

5 CONCLUSION

We propose a semi-supervised segmentation method using SupMix and SUFD, demonstrating superior results compared to conventional semi-supervised learning methods. However, since the accuracy for the most challenging pleural effusions class in the COVID-19 dataset did not improve, we plan to enhance the performance by considering prior probabilities and placing pleural effusions class labels in appropriate positions rather than simply pasting them.

ACKNLOWLEDGEMENTS

This paper is partially supported by the Strategic Innovation Creation Program.

REFERENCES

Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P. (2019). Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 622–637. Springer.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818.

Chen, W., Zhang, T., and Zhao, X. (2021a). Semantic segmentation using generative adversarial network. In *2021 40th Chinese Control Conference (CCC)*, pages 8492–8495.

Chen, X., Yuan, Y., Zeng, G., and Wang, J. (2021b). Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2613–2622.

Devries, T. and Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *ArXiv*, abs/1708.04552.

Fraz, M. M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A. R., Owen, C. G., and Barman, S. A. (2012). Chase db1: Retinal vessel reference dataset.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guo, J., Si, Z., Wang, Y., Liu, Q., Fan, M., Lou, J.-G., Yang, Z., and Liu, T. (2021). Chase: A large-scale and pragmatic chinese dataset for cross-database context-dependent text-to-sql. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2316–2331, Online. Association for Computational Linguistics.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jeong, J., Lee, S., Kim, J., and Kwak, N. (2019). Consistency-based semi-supervised learning for object detection. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Olsson, V., Tranheden, W., Pinto, J., and Svensson, L. (2021). Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1369–1378.

QMENTA (2020). Covid-19 ct segmentation dataset. `https://www.qmenta.com/blog/covid-19-ct-segmentation-dataset`. Accessed: 2024-09-26.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608.

Souly, N., Spampinato, C., and Shah, M. (2017). Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE international conference on computer vision*, pages 5688–5696.

Tsuda, H. and Hotta, K. (2019). Cell image segmentation by integrating pix2pixs for each class. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Yang, L., Qi, L., Feng, L., Zhang, W., and Shi, Y. (2023). Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032.

Zenati, H., Foo, C. S., Lecouat, B., Manek, G., and Chandrasekhar, V. R. (2018). Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. (2020). Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008.

Zou, Y., Zhang, Z., Zhang, H., Li, C.-L., Bian, X., Huang, J.-B., and Pfister, T. (2021). Pseudoseg: Designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations*.