M.S. THESIS

Voxel-wise adversarial semi-supervised learning for medical image segmentation

의료영상에서 다중 장기 분할을 위한 복셀 간 적대적 학습을 이용한 준 지도 학습 알고리즘

2023 년 2 월

서울대학교 대학원

컴퓨터 공학부

박 혜 림

M.S. THESIS

Voxel-wise adversarial semi-supervised learning for medical image segmentation

의료영상에서 다중 장기 분할을 위한 복셀 간 적대적 학습을 이용한 준 지도 학습 알고리즘

2023 년 2 월

서울대학교 대학원

컴퓨터 공학부

박 혜 림

Voxel-wise adversarial semi-supervised learning for medical image segmentation

의료영상에서 다중 장기 분할을 위한 복셀 간 적대적 학습을 이용한 준 지도 학습 알고리즘

지도교수 신 영 길

이 논문을 공학석사학위논문으로 제출함

2022 년 12 월

서울대학교 대학원

컴퓨터 공학부

박 혜 림

박혜림의 공학석사 학위논문을 인준함

2022 년 12 월

위 원 장	서 진 욱	(인)
부위원장	신 영 길	(인)
위 원	이 영 기	(인)

Abstract

Semi-supervised learning for medical image segmentation is an important area of research for alleviating the huge cost associated with the construction of reliable large-scale annotations in the medical domain. Recent semi-supervised approaches have demonstrated promising results by employing consistency regularization, pseudo-labeling techniques, and adversarial learning. These methods primarily attempt to learn the distribution of labeled and unlabeled data by enforcing consistency in the predictions or embedding context. However, previous approaches have focused only on local discrepancy minimization or context relations across single classes. In this paper, we introduce a novel method that effectively embeds both local and global features from multiple hidden layers and learns context relations between multiple classes. Our voxel-wise adversarial learning method utilizes a voxel-wise feature discriminator, which considers multilayer voxel-wise features as an input by embedding class-specific voxel-wise feature distribution. The experimental results demonstrate that our method outperforms current best-performing state-of-the-art semi-supervised learning approaches by improving the network performance by 2% in Dice score coefficient for multi-organ dataset. Furthermore, visual interpretation of the feature space demonstrates that our proposed method enables a well-distributed and separated feature space from both labeled and unlabeled data, which improves the overall prediction results.

Keywords: adversarial learning, feature discriminator, medical image segmentation, representation learning, semi-supervised learning

Student Number: 2021-20348

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Medical image segmentation is an essential task in several clinical approaches, such as computer-aided diagnosis, radiation therapy, and virtual surgeries [1, 2, 3]. Automated segmentation of organs (e.g., left atrium (LA), heart, or liver) is of significant importance in optimizing clinical workflow, such as the planning of surgeries and treatments. Convolutional neural networks (CNNs), which have demonstrated significant abilities in learning visual features in computer vision tasks, have been successfully adapted to medical segmentation problems by leveraging a large amount of annotated medical data (i.e., computed tomography (CT) scans [4]). However, the generation of reliable large-scale annotations of three-dimensional (3D) medical images requires domain-specific expertise, which is expensive and time-consuming.

Significant efforts, such as pretraining, self-supervised learning, and active learning, have been dedicated towards learning from a large number of unlabeled datasets. Semi-supervised learning is one of the approaches used to reduce the annotation cost, where the method simultaneously utilizes a large number of un-

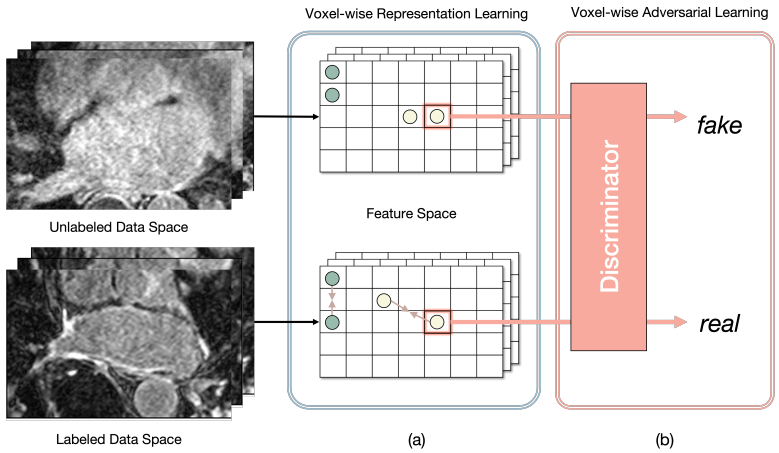


Figure 1.1: Our proposed method

labeled datasets with a limited number of labeled datasets. The semi-supervised approach assumes that labeled and unlabeled data from the same label share the same or similar underlying distribution (i.e., manifold assumption) [5, 6]. We can infer that labeled and unlabeled data usually share similar distributions (e.g., intensities or structures) in medical imaging; consequently, rich semantic information can be embedded using unlabeled data via semi-supervised learning. In practice, several studies on semi-supervised medical image segmentation has proposed effective methods for leveraging unlabeled data. Consistency regularization [7], pseudo-labeling [8] and adversarial learning [9] methods are some of the most commonly used learning methods in semi-supervised learning. The teacher-student model architecture [7] has been broadly applied, and it was demonstrated to be effective for consistency regularization- and pseudo-labeling-based methods. Furthermore, improved model performance can be expected through the synergy of representation learning methods from self-supervised and supervised learning by encoding representations from labeled data. Training with these representation learning methods [10, 11] showed rich

representation space especially for multi-class (i.e., multi-organ) dataset, which has not been explored much in semi-supervised learning.

Consistency regularization-based methods [7] learn network outputs that are invariant to perturbations or augmentations by adding noise to the unlabeled samples. Different types of methods have been presented to enforce consistency between outputs from different passes, such as uncertainty-aware schemes for data-level consistency [12] or task-level consistency using a task-transform layer [13]. Pseudo-labeling-based methods [14] generate high-confidence training targets as pseudo-labels for training unlabeled samples. Similar to consistency-based methods, the generated pseudo-labels are used to encourage mutual consistency [14] to enhance the generalized feature performance. However, these methods learn features by minimizing the loss function in the last layer (i.e., decision space), which can be limited to the local region so that the model learns only the local features of data.

On the other hand, adversarial learning-based methods [9, 15] model data distribution of unlabeled samples in an unsupervised setting by utilizing a discriminator. To capture the global shape constraint, a shape-aware adversarial learning method [15] has been proposed for unlabeled data. Although this method is effective for learning shape-aware global features, reproducing features through a separate network is ineffective for learning. Furthermore, both consistency- and adversarial learning- based methods only consider single-class cases and can be limited when they are extended to a multiclass dataset. Motivated by these recent studies, this work attempts to apply consistency-, pseudo-labeling-, and adversarial-based methods for leveraging unlabeled data. However, more effective learning methods are still significantly required for semi-supervised medical image segmentation task.

In this work, our goal is to improve the representation power for the med-

ical image segmentation task (similar to our previous work [11]), by leveraging a large amount of unlabeled data. Specifically, we intended to present an effective method that could successfully learn both local and global features from labeled and unlabeled datasets. However, there were several limitations associated with increasing representation power in previous studies. First, these studies focused only on local discrepancy minimization. Most consistency-based methods [12, 14, 16] calculate output discrepancy in the last layer such that only local features are embedded throughout the training scheme. However, both local and global features should be considered to obtain a better representation space. Second, feature relations across different classes of organs are ignored. Previous studies have only discussed the effectiveness of their methods for single-class data by embedding voxel-to-voxel local relations without distinguishing among different classes. The feature relation between different classes is also important for multiclass data.

This work proposes a novel adversarial learning-based method to incorporate unlabeled data to improve the network performance. We propose a context-aware semi-supervised segmentation method for efficiently learning the distributions of labeled and unlabeled datasets. To resolve the aforementioned problems of recent studies, we considered voxel-wise features from multiple hidden layers, which include both the local and global information of the data, as an input to our voxel-wise feature discriminator to embed distributions of unlabeled datasets.

Our proposed method is illustrated in Fig. 1.1. Existing semi-supervised segmentation models learn to map voxels from the data space to the feature space, ignoring global features or class-wise voxel relations. We enforced models to directly learn features representations of labeled and unlabeled data using our proposed method; by defining voxel-wise feature relations of labeled data in the

feature space as described in Fig. 1.1(a) (i.e., voxel-wise representation learning) and by discriminating between the voxel-level features from labeled and unlabeled data as presented in Fig. 1.1(b) (i.e., voxel-wise adversarial learning). As illustrated in Fig. 1.1b, the job of this discriminator is to determine if a voxel-wise feature belongs to labeled data or unlabeled data (real for labeled data and fake for unlabeled data). This voxel-wise feature discriminator assumes the form of a multitask discriminator that can learn distributions from different classes simultaneously, thereby allowing us to embed class-specific context-aware features in the embedding space.

Furthermore, we propose an improved voxel-wise representation learning method (Fig. 1.1a) for labeled data. To effectively embed unlabeled data, we are required to implement well-distributed features from labeled data prior to adversarial learning. In our previous study [11], we presented an explicit representation learning method for a supervised segmentation task by defining voxel-level feature relations. We adjusted this previous method to embed feature representations from labeled data without information loss using a multiresolution context resizing technique. Moreover, we used the Bootstrap Your Own Latent (BYOL) approach [17], instead of SimSiam [18], for learning stability.

To summarize, our contributions are as follows:

- We propose a **voxel-wise adversarial learning** method that learns both the local and global contexts of labeled and unlabeled data to avoid the local discrepancy problem from previous studies by considering voxel-wise features as input. Furthermore, our voxel-wise feature discriminator embeds feature relations across different classes by learning a class-specific voxel-wise feature distribution (instead of considering only a single class).

- We improve the previous **voxel-wise representation learning** method

by overcoming information loss and learning stability problems. This enables our adversarial learning method to effectively learn well-distributed voxel-wise feature representations.

- Our method achieves superior results on the Left Atrial Segmentation Challenge dataset and abdominal multiorgan (MO) dataset when compared with existing state-of-the-art semi-supervised segmentation methods (i.e., consistency regularization, pseudo-labeling and adversarial learning based methods).

Chapter 2

Related Works

2.1 Semi-Supervised Medical Image Segmentation

For semi-supervised medical image segmentation, traditional methods, such as prior- [19] and clustering-based models [20], use hand-crafted features to enhance model performance. With the advanced ability of CNNs, deep learning-based approaches have been widely used for medical image segmentation. Recently, semi-supervised methods based on consistency regularization [12, 13], pseudo labeling [14], and adversarial learning-based [21, 15, 22] have proven the effectiveness of incorporating a large amount of unlabeled data for medical image segmentation task.

Consistency Regularization. Consistency regularization is based on the assumption that the segmentation prediction of a network is consistent under realistic perturbations. This motivation was first proposed in [23] and further studied in [24, 7]. The Π-Model [24] encourages consistent training under different augmentation and dropout conditions. Owing to the noisy training tar-

get problem, temporal ensembling [24] adopts the exponential moving average (EMA) of previous evaluations to obtain an ensemble prediction. As a more time-effective method, the teacher-student model [7] introduces a pair of networks (i.e., teacher and student networks) and enforces consistency in their predictions. Time efficiency and accuracy can be achieved by averaging model weights, instead of label predictions.

In medical research, the uncertainty-aware mean teacher (UA-MT) model, proposed in [12], utilizes an uncertainty-aware teacher-student framework for LA segmentation. The base model framework was extended from the teacher-student architecture [7], and uncertainty map guidance was adopted to filter out unreliable predictions. More recently, a dual-task consistency (DTC) model [13] simultaneously used a pixel-wise segmentation map and level set representation as dual tasks. By utilizing the level set representation, the network could learn the geometric prior. However, the aforementioned methods tend to consider only the local context from the last layer, which can limit the representation of rich global contextual features in the embedding space.

Pseudo-labeling. The concept of pseudo-labeling was proposed in [8], and its variants have presented significant results in semi-supervised learning. For instance, NoisyStudent [25] employed a pair of networks, one acting as a teacher and the other as a student. They first trained the teacher network and inferred pseudo-labels for unlabeled images using the teacher network. A larger student network model was then trained using a combination of labeled and pseudo-labeled data, and this process was iterated by converting the student to the teacher. Moreover, a mutual consistency network (MC-Net) [14] proposed a cycled pseudo-label scheme that used one encoder and two marginally different decoders to utilize unlabeled data. Our method also adopts pseudo-labeling based on teacher-student architecture to infer voxel-wise features from unla-

beled data in a simple yet powerful manner.

Adversarial Learning. Inspired by the concept of generative adversarial networks (GANs) [26], several methods that use adversarial learning to exploit unlabeled data have attracted attention in semi-supervised medical image segmentation. For instance, [27, 28] used GANs to expand the training set to increase data diversity and avoid overfitting. Another key idea of using GANs in semi-supervised learning is to force the statistical prior-shape distribution and prediction distribution to be close so that they can effectively learn the distribution on the entire dataset (both labeled and unlabeled). A shape-aware semi-supervised segmentation network (SASSNET) [15] employs GANs to learn the distribution of both labeled and unlabeled data. This method utilizes the signed distance map (SDM) of images as an input to the discriminator, which plays a vital role in embedding the geometric context of unlabeled data. Although this method [15] considers global features employing SDM and a discriminator, context relations between different classes cannot be considered.

2.2 Representation Learning

Self-supervised learning methods [29, 30, 31] based on contrastive loss have proven to be effective in representation learning. Moreover, representation learning has also demonstrated its significance in semi-supervised medical image segmentation [32, 33]. In contrastive learning, positive (similar) pairs are pulled close together, whereas negative (dissimilar) pairs are pushed away. Because more negative samples can prevent collapse [29], several approaches, such as large batch sizes [31] or memory banks [30], have been proposed. Meanwhile, non-contrastive based approaches [17, 18] have shown effective results that avoid collapsing without using negative samples. The BYOL [17] method is based on

teacher-student model, and one branch of momentum encoder enables the network to learn representations without negative samples. Similarly, SimSiam [18] uses a Siamese network and stop-gradient operation, instead of momentum encoder, to prevent collapsing.

These non-contrastive based approaches can be employed in supervised learning to learn rich representations [11]. Inspired by SimSiam [18], our previous study [11] presented an effective representation learning method for medical segmentation task by defining voxel-level relations in the embedding space. In this study, we improved our previous method by solving the information loss and learning instability problems of Siamese networks.

Chapter 3

Proposed Method

We aim to learn feature representation (i.e., local and global features) from both the labeled and unlabeled datasets. To achieve this, we propose a context-aware semi-supervised segmentation method that can be incorporated into a segmentation network (i.e., VNet [34]). The overall architecture of semi-supervised segmentation is illustrated in Fig. 3.1. Two backbone networks (i.e., VNet [34]), i.e., teacher and student networks, take computed tomography scans as an input. The teacher network is learned passively via exponential mean average (EMA). The features ($\bigcirc$) from multiple hidden layers of the student network pass through each section of our proposed network (i.e., voxel-wise feature layer and voxel-wise feature discriminator) so that feature representations from labeled and unlabeled data can be learned. The features ($\triangle$) of the teacher network are used for optimizing the student and our proposed network. The student network is trained using four loss functions (L_{dice}, L_{c}, L_{adv}, and $\mathrm{L}_{feature}$). The gradients are not backpropagated through the dashed lines.

We assume a set of training sets containing N labeled data and M unla-

beled data, where $N \ll M$. We denote the labeled set as $\mathcal{D}_l = \{(x_i, y^i_{gt})\}^N_{i=1}$ and unlabeled set as $\mathcal{D}_u = \{(x_i)\}^{N+M}_{i=N+1}$, where x_i represents the 3D volume, and y^i_{gt} denotes the ground-truth label. The proposed architecture for semi-supervised learning consists of two parts: voxel-wise representation learning (the blue box in Fig. 3.1) and voxel-wise adversarial learning (the red box in Fig. 3.1). Features from the hidden layers of the backbone network pass through each part to learn feature representations from $\mathcal{D}_l$ and $\mathcal{D}_u$. The voxel-wise adversarial learning method takes voxel-wise features from $\mathcal{D}_l$ and $\mathcal{D}_u$, after which it learns class-specific data distributions. The voxel-wise representation learning method uses voxel-wise features from $\mathcal{D}_l$ and improves current embeddings by defining feature relations from the same class. In Section 3.1 and 3.2, we describe the details of these methods. In Section 3.3, we explain the overall training process of our proposed method.

3.1 Voxel-wise Adversarial Learning

To leverage a large amount of unlabeled data, the network must be able to learn feature representations using only CT images. Previous consistency-based methods [12, 13] have applied a consistency loss function and trained the network for consistent prediction with perturbed or transformed outputs. The consistency loss is computed between y_{pseudo} and y for labeled and unlabeled data. However, this loss is computed in the last layer (i.e., decision space), which embeds only the local features of data. Moreover, it penalizes voxel-wise consistency ignoring class-specific information. It is also problem in [15] that embedded shape-aware global features are only limited to a single class.

To resolve this problem, we propose a novel voxel-wise feature discriminator for embedding class-specific features of both labeled and unlabeled data.

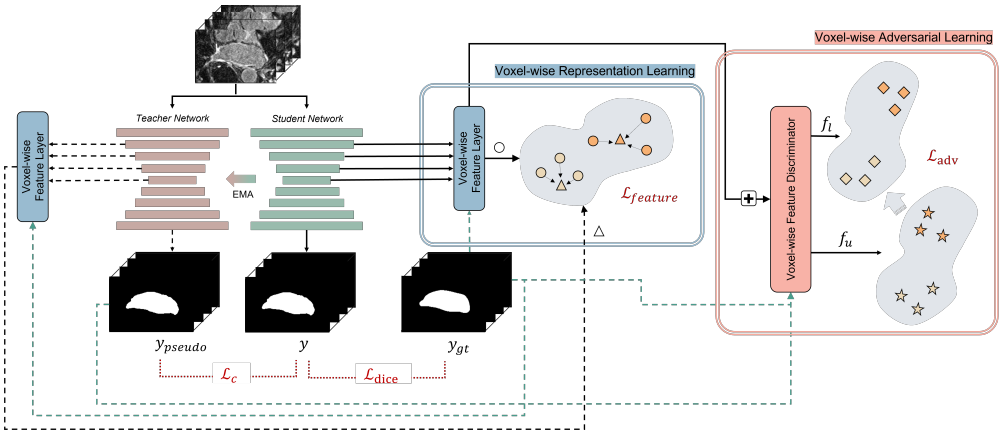


Figure 3.1: Overview of the proposed architecture

As presented in Fig. 3.2a, our voxel-wise feature discriminator takes a set of multiresolution features, $\{E(x^1), E(x^2), E(x^3), E(x^4)\}$, as an input, where $E(\cdot)$ denotes an encoder of the backbone, and $E(x^j) \in R^{H \times W \times D \times C}$ denotes features from the j^{th} hidden layer. These features from multiple hidden layers pass through the convolution layer to adjust the channel size, and each feature is upsampled to the same spatial size. Such features from multiple hidden layers are fused into one by adding an operation and a convolution layer. Thereafter, voxel-level features (C-d vector) from this fused feature, f_{fused}, pass through a voxel-level feature discriminator, which consists of two multilayer perceptron networks (MLPs) and prediction layer (i.e., linear branch). The number of prediction layers corresponds to the number of class (in case of LA dataset, there exist two classes; foreground and background). The voxel-level features from different classes pass through different prediction layers. To specify the class of each voxel-level feature, we use ground-truth label y_{gt} for labeled data and pseudo-labels y_{pseudo} for unlabeled data, which can be computed using the

following equation:

$$y_{pseudo} = \text{argmax } \{Teacher(x) > t\}, \tag{3.1}$$

where t represents the threshold parameter, which lies in the range of $[0, 1]$.

This different prediction branches enable multiple simultaneous adversarial classification tasks. We define features from labeled data as real and those from unlabeled data as fake so that the encoder of the segmentation network (generator) can generate voxel-level features of unlabeled data with a distribution similar to that of voxel-level features of labeled data. This forces the distributions of class-specific voxel-level features from both labeled and unlabeled features to be close. In this manner, the segmentation network can learn class-specific context-aware features more effectively. The encoder can embed both local and global features using a multiresolution context-fusion technique. In D representing the voxel-wise feature discriminator, we can define our proposed adversarial loss function as follows:

$$\begin{aligned} \mathrm{L}_{adv} = \frac{1}{N} \sum_{n=1}^{N} \sum_{f_i \in E(x_n)} log(D(f_i)) + \\ \frac{1}{M} \sum_{m=N+1}^{N+M} \sum_{f_i \in E(x_m)} log(1 - D(f_j)). \end{aligned} \tag{3.2}$$

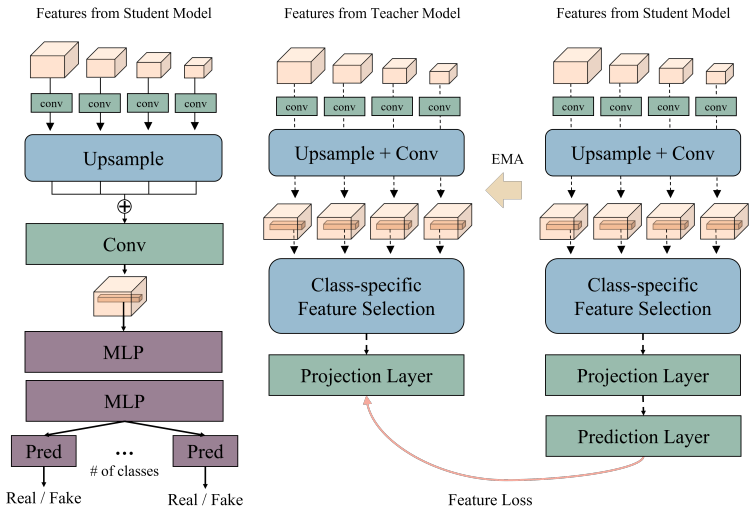
3.2 Voxel-wise Representation Learning

In Section 3.1, we propose a new voxel-wise feature discriminator for learning the feature representations of unlabeled data via learning based on the feature distribution of labeled data. In this setting, the most important task is the modeling of the distribution of features from labeled data beforehand. Accurate modeling of the labeled data distribution is essential for effective adversarial learning. In other words, the model is unlikely to learn effectively from

adversarial learning if the distribution of labeled data is incorrect. In contrast, the model is likely to learn effectively if distribution is recovered from labeled data. Thus, our model can learn rich feature representations from both labeled and unlabeled data.

In our previous work [11], a voxel-level Siamese representation learning method for medical image segmentation tasks was proposed. By defining voxel-wise feature relations in the representation space, the model learned feature representations that were effective in the segmentation task. The stop-gradient technique and the Siamese network from SimSiam [18] were used to learn voxel-wise feature relations. A feature aggregation method was also proposed for embedding both local and global features. However, our previous study had two limitations: (1) learning stability and (2) information loss.

In this study, we propose an improved voxel-wise representation learning method for embedding features from labeled data. Inspired by previous studies [17, 35], we used the learning technique from BYOL [17], instead of SimSiam [18], for the first problem(i.e., learning stability). Using EMA from BYOL enabled the model to produce a more stable prediction target [35] than the stop-gradient technique from SimSiam [18]. As presented in Fig. 3.2b, there are teacher and student models; however, the teacher model uses the slow moving average of the student parameter, instead of learning for its own parameter (i.e., EMA). The weights of the teacher θ_t are updated as $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$, where λ represents the decay parameter, and θ_s indicates the weights of the student. Furthermore, for the second problem (i.e., information loss), multiresolution context resizing method is proposed. The information loss occurs during the downsampling of mask data to match the class location for each voxel-wise feature. Thus, instead of downsampling the mask data, the multiresolution features from the encoder, $E(\cdot)$ were upsampled. Figure 3.2b illustrates the upsampling



(a) Voxel-wise Feature Discriminator

(b) Voxel-wise Feature Layer

Figure 3.2: Details of the proposed architecture.

and convolution stage that can reduce information loss.

As explained in Section 3.1, our voxel-wise feature layer (Fig. 3.1 and Fig. 3.2b) uses multiresolution features from the encoder of the backbone as an input. These features pass through the upsampling and convolution stages, and voxel-wise features, p_i^c, are selected for each class; here, p_i^c refers to the i^{th} voxel-wise feature from class c (class-specific feature selection). These sampled voxel-wise features pass through the projection and prediction layers. The projection layer from the teacher network outputs z_t, and the projection and prediction layers from the student network output $p(z_s)$, where $p(\cdot)$ denotes the prediction layer. Based on a previous research [17], we used the mean square error between normalized z_t and $p(z_s)$ as the feature loss function. The feature loss function

for updating the student network can be defined as follows:

$$\mathrm{L}_{feature} = \|\bar{p}(z_s) - \bar{z}_t\|_2^2, \tag{3.3}$$

where $\bar{x}$ refers to l_2-normalization (i.e., $\bar{x}\frac{x}{\|x\|_2}$).

3.3 Training Details

Our backbone network is based on VNet [34]. We first demonstrate a basic VNet [34] segmentation training scheme for a labeled dataset. Two VNets [34] are displayed in Fig. 3.1: the teacher and student networks. These two networks take the 3D volume, $x \in R^{H \times W \times D}$, as an input, and they output prediction masks, y_{pseudo} and y respectively. Based on [12, 15], we used the dice loss [36] to maximize the overlap between the ground truth and prediction y to train the student network. We used the labeled dataset (i.e., $\mathcal{D}_l$) to compute the dice loss, which can be defined as

$$\mathrm{L}_{dice} = \sum_{i=1}^{N} \frac{2y_i \cdot Student(x_i)}{(y_i)^2 + (Student(x_i))^2}. \tag{3.4}$$

For updating the teacher network, we used the EMA [35] technique.

Following [7], we also added a consistency loss between the softmax predictions of the teacher and student networks for semi-supervised learning. The consistency loss between the outputs of the teacher and student networks can be summarized as follows:

$$\mathrm{L}_c = E_x \left[\|f(x, \theta_t) - f(x, \theta_s)\|^2 \right], \tag{3.5}$$

where $f(\cdot)$ represents the VNet architecture [34]. We can stabilize the label prediction by using the teacher-student framework and penalize the predictions that are inconsistent with the target (i.e., output of the teacher network) by

adding consistency loss. In this manner, we can learn the generalized local features of both labeled and unlabeled datasets.

The final loss function for training the student network (i.e., VNet [34]) is summarized as follows:

$$\mathrm{L}_{total} = \alpha \mathrm{L}_{adv} + \beta \mathrm{L}_{feature} + \gamma \mathrm{L}_{c} + \mathrm{L}_{dice}, \tag{3.6}$$

where α, β and γ represent the coefficients used to balance the different loss terms.

3.4 Dataset Details

We evaluated our method using two datasets: the LA dataset from the Atrial Segmentation Challenge and an MO dataset.

Left Atrial Segmentation Challenge dataset We used 100 3D gadolinium-enhanced magnetic resonance imaging scans and an LA segmentation mask for training and validation. In the dataset, the scans exhibited an isotropic resolution of $0.6255mm^3 \times 0.6255mm^3 \times 0.625mm^3$. Following the settings of a previous method [12, 15], the dataset was separated into two sets: training and testing, with 80 images for training and 20 for testing. We applied the same preprocessing method; we randomly cropped $112 \times 112 \times 80$ sub-volumes and preprocessed the scans using a windowing range of $[-125, 275]$. Then, we normalized the input images to zero mean and unit variance (i.e., the range of the value is $[0, 1]$). We used data augmentations techniques; randomly flipping and rotating.

Abdominal multiorgan dataset To further evaluate the effectiveness of our method in multiclass segmentation, we evaluated its performance on an MO dataset. We used 90 abdominal CT images: 47 from the Beyond the Cranial Vault dataset [37] and 43 from the Pancreas-CT dataset. The segmenta-

tion standard consisted of the spleen, left kidney, gallbladder, esophagus, liver, stomach, pancreas, and duodenum. The slice thickness was in the range of $0.5 - 5.0mm$ and pixel sizes were in the range of $0.6 - 1.0mm$. The dataset was separated into two sets: 70 images for training and 20 for testing. We sampled all abdominal CT images into $128 \times 128 \times 64$ pixels. Preprocessing was performed using a soft-tissue CT windowing range of $[-200, 250]$ Hounsfield units. After rescaling, we normalized the input images to zero mean and unit variance (i.e., the range of the value is $[0, 1]$).

3.5 Implementation Details

For training both LA and MO dataset, we used a VNet [34] architecture as the base network. We set the batch size to 4, and each batch included two labeled patches and two unlabeled patches.

For the LA dataset, we used the stochastic gradient descent optimizer (momentum = 0.9, weight decay of 0.0001) for 6000 iterations, with an initial learning rate of 0.01. The learning rate was divided by 10 for every 2500 iterations. To train the multitask feature discriminator, we followed the method described in [38]; we used an Adam optimizer (β_1=0.5, β_2=0.999) and a learning rate of 0.0002. The weighting parameter α was 0.01 for L_{adv} and β was 0.1 for $L_{feature}$. Following [15], we used Gaussian warming-up function $\gamma(t) = 0.001 * e^{-5(1-\frac{t}{t_{max}})^2}$ for consistency loss where t indicates the number of iterations. Based on our previous study[11], the dimensions of all hidden layers from in voxel-level feature layer were set to 64. Furthermore, we used a threshold t of 0.7. We implemented our framework in PyTorch [39], using an NVIDIA TITAN RTX GPU and Tesla V100 GPU. At the inference time, only the VNet framework was used for segmentation.

For MO dataset, we used Adam optimizer (β_1=0.9, β_2=0.999) and an initial learning rate of 0.001 decayed by 0.1 every 2500 iterations. The weighting parameter α was 0.01 for L_{adv} and β was 100 for $L_{feature}$. The rest of the experimental settings were the same as those employed in the LA dataset experiments.

3.6 Evaluation Metrics

For our evaluation metrics, we determined the dice score coefficient (DSC) [36], Hausdorff distance (HD95; mm) [40, 41], average symmetric surface distance (ASSD; mm) [42], and Jaccard Index (JC). Given the binary labeled masks X and Y, the DSC and JI are defined as follows:

$$DSC(X,Y) = \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{3.7}$$

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \tag{3.8}$$

As a better-generalized evaluation metric for distance, HD is defined as follows:

$$HD(X,Y) = max\{\underset{s_X \in S_X}{max}\, d(s_X, S_Y) + \underset{s_Y \in S_Y}{max}\, d(s_Y, S_X)\} \tag{3.9}$$

where S_X is a set of surface voxels of a set X, and $d(p, S_X)$ is the shortest distance from an arbitrary voxel p to S_X

$$d(p, S_X) = \underset{s_X \in S_X}{min}\, \|p - s_X\|_2 \tag{3.10}$$

By Defining the distance function as

$$D(S_X, S_Y) = \sum_{s_X \in S_X} d(s_X, S_Y) \tag{3.11}$$

Therefore, the ASSD is defined as follows:

$$ASSD(X,Y) = \frac{1}{|S_X| + |S_Y|}(D(S_X, S_Y) + D(S_Y, S_X)) \tag{3.12}$$

Chapter 4

Experimental Results

4.1 Results

Left Atrial Segmentation Challenge dataset. We evaluated the performance of our proposed network in terms of its accuracy by comparing our results with those of the state-of-the-art models, i.e., domain-agnostic prior [43], UA-MT [12], SASSNet [15], local and global structure-aware entropy regularized mean teacher [44], double-uncertainty weighted method [45], DTC [13], contrastive voxel-wise representation learning [32], and MC-Net [14]. Two semi-supervised settings widely used on the LA dataset were available from a previous study [15] (i.e., using either 10 or 20% of the labeled data). Table 4.1 lists the quantitative results of LA segmentation. The results indicate that our proposed method achieves superior results in terms of the DSC, Jaccard index, and HD95 measurements and achieves competitive results on ASSD under the conditions of both 10% and 20% labeled data. Qualitative results are illustrated in Fig. 4.2. It can be observed that our method has a higher overlap ratio with respect

to the ground truth in both 2D and 3D visualizations, thereby producing fewer false positives.

Abdominal multi-organ dataset. To prove the effectiveness of our method on a multiclass dataset, we conducted an experiment on an MO dataset. For comparison, several state-of-the-art models (i.e., UA-MT [12], SASSNet [15], DTC [13], and MC-Net [14]) and the base network, VNet, were used for evaluation. We considered 20% of training data among the 70 images as the labeled data (14 labeled) and the others as unlabeled data (54 unlabeled). All the models used VNet as their backbone network. Table 4.2 presents quantitative comparisons of the segmentation results.

The results indicate that our method outperforms the other methods in terms of all evaluation metrics (i.e., Dice (71.28%), Jaccard index (59.01%), HD (4.32), and ASSD (1.24)). Our method achieves significant improvements in the segmentation of spleen, liver, stomach, and pancreas and demonstrates competitive results for other organs. A box plot for a more precise quantitative comparison is presented in Fig. 4.1. The qualitative results illustrated in Fig. 4.3 indicate that our method segments multiple organs better than other methods. The statistical significance of all the experiments was verified through paired sample t-tests (i.e., p-values) with the proposed method based on a 95% confidence interval. The p-values that were less than 0.03 were considered to be statistically significant.

4.2 Ablation Study

We performed an ablation study to investigate the effectiveness of major components of the proposed loss function. We trained VNet under 20% labeled data using the MO and LA datasets, and the results are listed in Table 4.3 and

Method	# Scans used		Metrics			
	Labeled	Unlabeled	Dice(%)	Jaccard(%)	95HD(voxel)	ASSD(voxel)
VNet	8(10%)	72	79.99	58.12	21.11	5.48
VNet	16(20%)	64	86.03	76.06	14.26	3.51
VNet	80(All)	0	91.14	83.82	5.75	1.52
DAP [43]	8(10%)	72	81.89	71.23	15.81	3.80
UA-MT [12]	8(10%)	72	84.25	73.48	13.84	3.36
SASSNet [15]	8(10%)	72	87.32	77.72	9.62	2.55
LG-ER-MT [44]	8(10%)	72	85.54	75.12	13.29	3.77
DUWM [45]	8(10%)	72	85.91	75.75	12.67	3.31
DTC [13]	8(10%)	72	86.57	76.55	14.47	3.74
CVRL [32]	8(10%)	72	87.72	78.29	9.34	2.23
MC-Net [14]	8(10%)	72	87.71	78.31	9.36	**2.18**
Ours	8(10%)	72	**88.42**	**79.38**	**8.74**	2.52
DAP [43]	16(20%)	64	87.89	78.72	9.29	2.74
UA-MT [12]	16(20%)	64	88.88	80.21	7.32	2.26
SASSNet [15]	16(20%)	64	89.54	81.24	8.24	2.20
LG-ER-MT [44]	16(20%)	64	89.62	81.31	7.16	2.06
DUWM [45]	16(20%)	64	89.65	81.35	7.04	2.03
DTC [13]	16(20%)	64	89.42	80.98	7.32	2.10
CVRL [32]	16(20%)	64	89.87	81.65	6.96	**1.72**
MC-Net [14]	16(20%)	64	90.34	82.48	6.00	1.77
Ours	16(20%)	64	**90.56**	**82.84**	**5.95**	1.79

Table 4.1: Quantitative comparisons on the left atrium dataset.

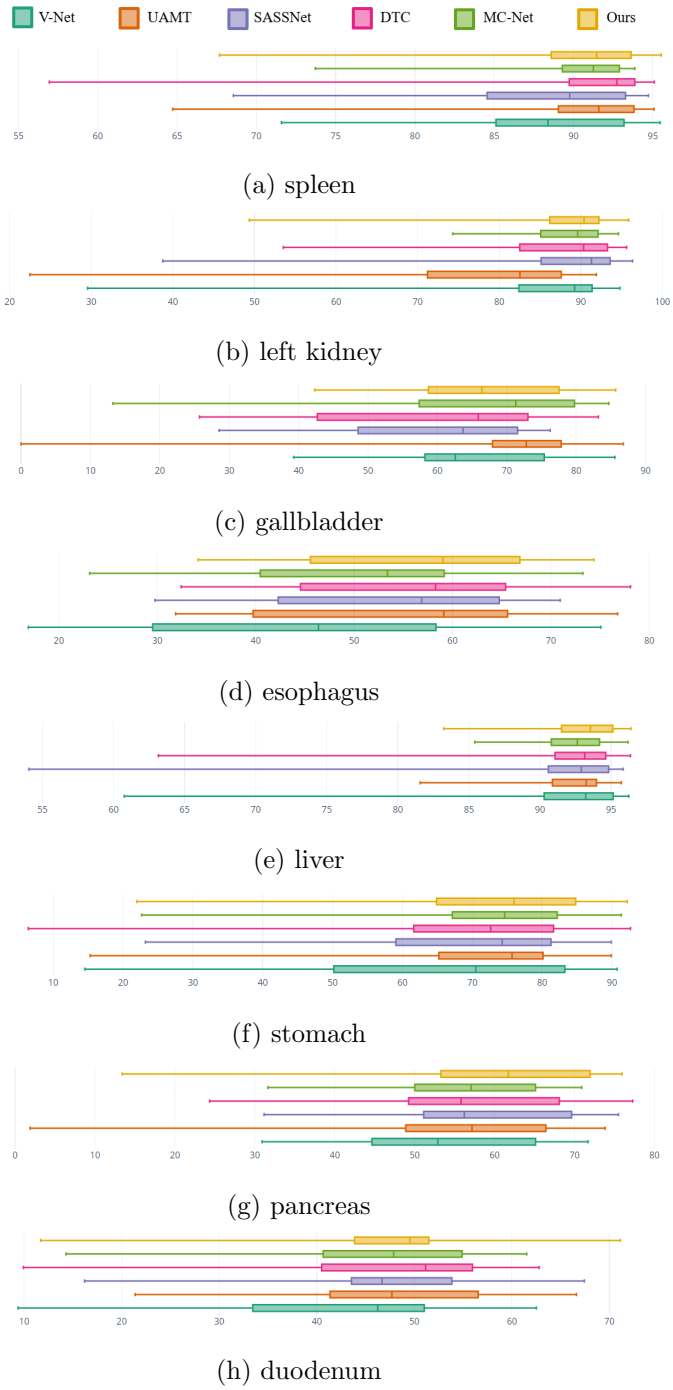


Figure 4.1: Box plots of the dice score coefficient of different methods for eight different organs.

Method	Metrics (average)				DSC							
	DSC(%)	JC(%)	HD(voxel)	ASSD(voxel)	spleen	left kidney	gallbladder	esophagus	liver	stomach	pancreas	duodenum
VNet [34]	66.58	54.08	5.74	1.79	87.79	81.98	64.69	44.88	91.00	66.51	53.39	42.42
UA-MT [12]	69.57	56.90	4.99	1.36	89.64	77.53	**67.82**	**56.19**	92.21	70.73	54.58	**47.86**
SASSNet [15]	69.09	56.42	4.85	1.48	87.42	87.26	60.19	54.16	90.41	69.41	57.30	46.59
DTC [13]	69.39	57.00	5.78	1.79	89.05	87.03	59.64	56.11	91.23	68.45	56.63	46.99
MC-Net [14]	69.76	57.34	5.61	1.90	89.15	**87.82**	64.66	50.50	92.28	71.22	56.97	45.49
Ours	**71.28**	**59.01**	**4.32**	**1.24**	**89.75**	87.07	66.64	56.01	**93.03**	**71.58**	**59.08**	47.03

Table 4.2: Quantitative comparisons on the multiorgan dataset

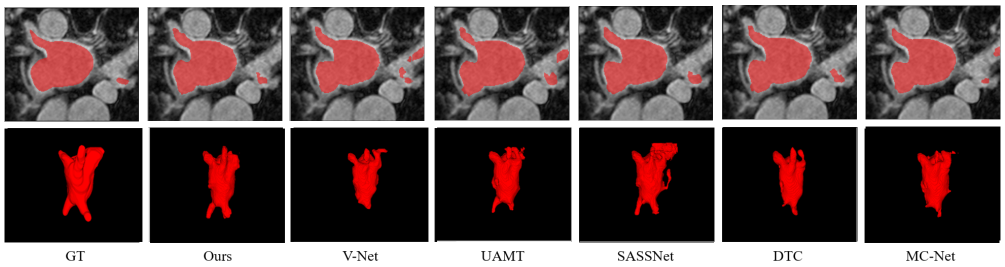


Figure 4.2: Qualitative comparison of different semi-supervised segmentation methods using the left atrium dataset with 20% labeled data.

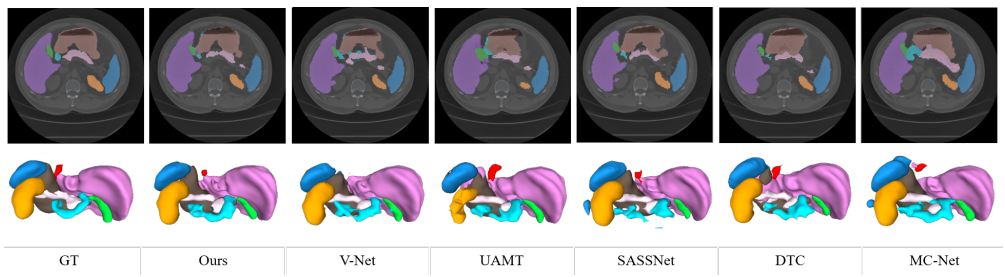


Figure 4.3: Qualitative comparison of different semi-supervised segmentation methods using the multiorgan dataset with 20% labeled data.

4.4, respectively. In Table 4.3, our network and MC-Net are trained with 20% labeled data using the MO datasets. We generated a visualization using labeled (marked by triangles) and unlabeled (marked by circles) data, and we present them separately below for comparison. Features are colored according to class

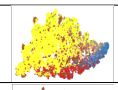
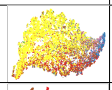
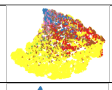
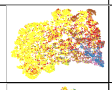
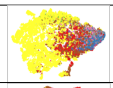
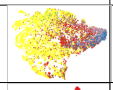
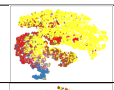
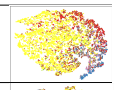
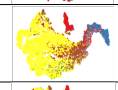
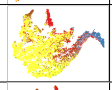
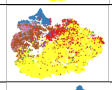
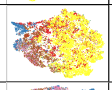
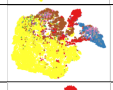
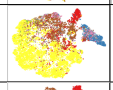
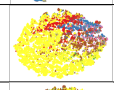
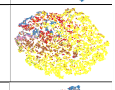
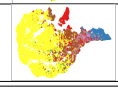
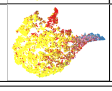
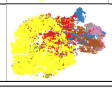
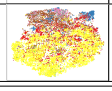
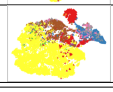
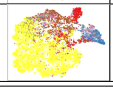
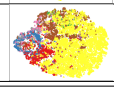
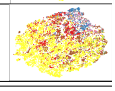
Iter	VNet + L_{adv}		VNet + $L_{feature}$		Ours (VNet + L_{total})		MC-Net[14]	
	labeled	unlabeled	labeled	unlabeled	labeled	unlabeled	labeled	unlabeled
0.1 k								
0.5 k								
1 k								

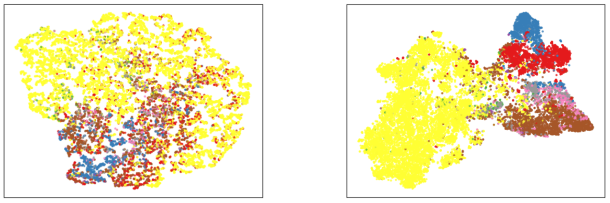
Table 4.3: Visualization of the feature alignment progress during the training phase.

Method	# Scans used		Metrics			
	Labeled	Unlabeled	Dice(%)	Jaccard(%)	95HD(voxel)	ASSD(voxel)
VNet	16(20%)	64	86.03	76.06	14.26	3.51
VNet+ L_{adv}	16(20%)	64	88.76	80.01	10.46	2.64
VNet+$L_{feature}$	16(20%)	64	88.67	79.85	11.52	3.31
VNet+L_{adv}+$L_{feature}$	16(20%)	64	90.39	82.56	10.11	2.70
VNet+L_{adv}+$L_{feature}$+L_c	16(20%)	64	**90.56**	**82.84**	**5.95**	**1.79**

Table 4.4: Ablation study of the effectiveness of our proposed method on the left atrium dataset

labels (the labels of unlabeled data are used for only visualization).

From Table 4.3, we can observe that each major component of our proposed method (i.e., L_{adv} and $L_{feature}$) contributes to a more structured representation space in the training process. Specifically, L_{adv} guides unlabeled data to follow the distribution of labeled data, and $L_{feature}$ plays a significant role in generating separated feature representation; this implies that the unlabeled data distribution follows the labeled data distribution as we intended, thereby embedding rich feature representation.



(a) VNet [34] (b) Ours

Figure 4.4: Visualization of features from the second layer.

Table 4.4 lists the comparison results of the ablations, wherein our losses (L_{adv}, $L_{feature}$, and L_c) were gradually incorporated. The results reveal a significant performance improvement in cases wherein the two losses, L_{adv} and $L_{feature}$, were used together, rather than being used separately. This demonstrates that these losses achieve synergy by learning the distribution of unlabeled features from well-distributed labeled features. Furthermore, including the loss function, L_c, achieves further improvements by stabilizing label prediction. The comparative analysis in terms of the four categories (e.g., single-class relation, multi-class relation, adversarial learning, and representation learning) is listed in Table 5.1. Our method can be applied to not only single-class relation, but also multi-class relation, and two main techniques (i.e., adversarial learning and representation learning) are successfully used for leveraging unlabeled data for training.

Chapter 5

Discussion

Recent semi-supervised segmentation approaches in medical imaging have demonstrated promising results by employing various techniques, such as consistency regularization [12, 13], pseudo-labeling [14], and adversarial learning [15]. However, previous methods train the network with the outputs obtained from the final layer, which complicates learning of global features by the network. Moreover, relations between different classes cannot be considered.

Our proposed method is effective for learning both local and global contexts by embedding voxel-level features with voxel-level feature layers and voxel-level feature discriminators (Table 4.1 and Fig. 4.2). We achieved a more structured representation space (Fig. 4.4 b) by defining voxel-level feature (including global and local context) relations in the representation space. The features are colored based on the class labels, and we visualize them using the test dataset (labels are only used for visualization). On comparison with a previous method [15] which also included global contextual information with the discriminator and SDM, our method achieved superior results (Table 4.2), particularly for multiclass

Categories	Single-class relation	Multi-class relation	Adversarial Learning	Representation Learning
UAMT [12]	✓	✗	✗	✗
SASSNet [15]	✓	✗	✓	✗
DTC [13]	✓	✗	✗	✗
CVRL [32]	✓	✗	✗	✓
MC-Net [14]	✓	✗	✗	✗
Ours	✓	✓	✓	✓

Table 5.1: Comparative analysis with the previous method in terms of four categories in the semi-supervised learning.

datasets. By learning class-specific voxel-level features using BYOL [17] and a multitask discriminator, we achieved separated representation space (Fig. 4.4 and Table 4.3) and precise segmentation results for the multiclass dataset (Table 4.2 and Fig. 4.3). This indicates that our method is effective for learning feature relations across different classes.

In future studies, we can improve the results by suggesting a more efficient method to enable unlabeled data to follow the distribution of labeled data. Furthermore, our method can be enhanced by developing more efficient feature sampling approaches.

Chapter 6

Conclusion

In this work, we propose a novel semi-supervised learning method for medical image segmentation tasks. Specifically, our voxel-wise representation learning method embedded feature representations (i.e., local and global features) in the representation space, and our voxel-wise feature discriminator successfully leveraged unlabeled data using the distribution of features from the labeled data.

Furthermore, our method could provide a more informative representation that embedded class-specific features and achieved superior results in multi-class segmentation. The experimental results demonstrated that our proposed method is specialized for embedding rich information from both labeled and unlabeled data, which brings additional improvement for medical image segmentation task. We believe that our approach can provide a useful perspective on medical imaging tasks and can be applied to various medical datasets, regardless of the number of classes.

Bibliography

[1] G. Glombitza, H. Evers, S. Hassfeld, U. Engelmann, and H.-P. Meinzer, "Virtual surgery in a (tele-)radiology framework," *IEEE Transactions on Information Technology in Biomedicine*, vol. 3, no. 3, pp. 186–196, 1999.

[2] R. D. Howe and Y. Matsuoka, "Robotics for surgery," *Annual Review of Biomedical Engineering*, vol. 1, no. 1, pp. 211–240, 1999. PMID: 11701488.

[3] van Ginneken B, S.-P. CM, and P. M, "Computer-aided diagnosis: how to move from the laboratory to the clinic.," *Radiology*, vol. 261, no. 3, pp. 719–732, 2011.

[4] M. Kudo, R. Q. Zheng, S. R. Kim, Y. Okabe, Y. Osaki, H. Iijima, T. Itani, H. Kasugai, M. Kanematsu, K. Ito, *et al.*, "Diagnostic accuracy of imaging for liver cirrhosis compared to histologically proven liver cirrhosis," *Intervirology*, vol. 51, no. Suppl. 1, pp. 17–26, 2008.

[5] O. Chapelle, B. Scholkopf, and A. Zien, Eds., "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[6] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, pp. 373–440, 2019.

[7] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2018.

[8] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[9] W. Li, Z. Wang, J. Li, J. Polson, W. Speier, and C. W. Arnold, "Semi-supervised learning based on generative adversarial network: a comparison between good gan and bad gan approach.," in *CVPR Workshops*, pp. 1–11, 2019.

[10] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12546–12558, 2020.

[11] C. E. Lee, M. Chung, and Y.-G. Shin, "Voxel-level siamese representation learning for abdominal multi-organ segmentation," *Computer Methods and Programs in Biomedicine*, vol. 213, p. 106547, 2022.

[12] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," 2019.

[13] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," 2021.

[14] Y. Wu, M. Xu, Z. Ge, J. Cai, and L. Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," *CoRR*, vol. abs/2103.02911, 2021.

[15] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3d semantic segmentation for medical images," *Lecture Notes in Computer Science*, p. 552–561, 2020.

[16] S. Li, Z. Zhao, K. Xu, Z. Zeng, and C. Guan, "Hierarchical consistency regularized mean teacher for semi-supervised 3d left atrium segmentation," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine &amp Biology Society (EMBC)*, IEEE, nov 2021.

[17] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284, 2020.

[18] X. Chen and K. He, "Exploring simple siamese representation learning," 2020.

[19] X. You, Q. Peng, Y. Yuan, Y.-m. Cheung, and J. Lei, "Segmentation of retinal blood vessels using the radial projection and semi-supervised approach," *Pattern recognition*, vol. 44, no. 10-11, pp. 2314–2324, 2011.

[20] N. M. Portela, G. D. Cavalcanti, and T. I. Ren, "Semi-supervised clustering for mr brain image segmentation," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1492–1497, 2014.

[21] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation uti-

lizing unannotated images," in *International conference on medical image computing and computer-assisted intervention*, pp. 408–416, Springer, 2017.

[22] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," *arXiv preprint arXiv:1802.07934*, 2018.

[23] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," *Advances in neural information processing systems*, vol. 27, 2014.

[24] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2017.

[25] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," 2020.

[26] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[27] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 289–293, 2018.

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

[29] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European conference on computer vision*, pp. 776–794, Springer, 2020.

[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[32] C. You, R. Zhao, L. Staib, and J. S. Duncan, "Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation," 2021.

[33] Y. Liu, W. Wang, G. Luo, K. Wang, and S. Li, "A contrastive consistency semi-supervised left atrium segmentation model," *Computerized Medical Imaging and Graphics*, vol. 99, p. 102092, 2022.

[34] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *CoRR*, vol. abs/1606.04797, 2016.

[35] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[36] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly un-

balanced segmentations," *Lecture Notes in Computer Science*, p. 240–248, 2017.

[37] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Automatic multi-organ segmentation on abdominal ct with dense v-networks," *IEEE Transactions on Medical Imaging*, vol. 37, no. 8, pp. 1822–1834, 2018.

[38] K. Kurach, M. Lucic, X. Zhai, M. Michalski, and S. Gelly, "A large-scale study on regularization and normalization in gans," 2019.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019.

[40] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[41] M. Chung, J. Lee, M. Lee, J. Lee, and Y.-G. Shin, "Deeply self-supervised contour embedded neural network applied to liver segmentation," *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105447, Aug 2020.

[42] F. Lu, F. Wu, P. Hu, Z. Peng, and D. Kong, "Automatic 3d liver location and segmentation via convolutional neural networks and graph cut," 2016.

[43] H. Zheng, L. Lin, H. Hu, Q. Zhang, Q. Chen, Y. Iwamoto, X. Han, Y.-W. Chen, R. Tong, and J. Wu, "Semi-supervised segmentation of liver using

adversarial learning with deep atlas prior," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 148–156, Springer, 2019.

[44] W. Hang, W. Feng, S. Liang, L. Yu, Q. Wang, K.-S. Choi, and J. Qin, "Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 562–571, Springer, 2020.

[45] Y. Wang, Y. Zhang, J. Tian, C. Zhong, Z. Shi, Y. Zhang, and Z. He, "Double-uncertainty weighted method for semi-supervised learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 542–551, Springer, 2020.

초록

의료 영상 분할에서 신뢰성 있는 큰 규모의 정답 레이블을 생성하기 힘들다는 점에서 준지도 학습은 중요하다. 최근 준 지도 학습 연구들은 일관성 정규화(consistency regularization), 의사 레이블링(pseudo-labeling) 그리고 적대적 학습(adversarial learning) 방법을 이용하여 좋은 결과를 보여주었다. 해당 방법들은 주로 정답이 있는 데이터(labeled data)와 정답이 없는 데이터(unlabeled data)의 분포를 학습하여 그들 간의 예측 값이나 임베딩 값에 일관성을 부여한다. 하지만 이전 방법들은 영상의 지엽적인 특성과 하나의 클래스 간의 관계만을 고려한다는 단점이 있다. 따라서 우리는 이 논문에서 영상의 지엽적인 특징과 전역적인 특징을 동시에 고려하며, 여러 클래스 간의 표현 관계를 학습할 수 있는 적대적 학습 기반의 준 지도 학습 의료 영상 분할 네트워크를 제안한다. 우리의 복셀 간 적대적 학습(voxel-wise adversarial learning) 메소드는 멀티레이어로부터 추출한 클래스 별 복셀 간 표현(voxel-wise feature)를 인풋으로 취급하는 복셀 간 표현 분류자(voxel-wise feature discriminator)를 활용한다. Left Atrial Segmentation Challenge data과 Abdominal Multi-Organ dataset을 이용한 실험을 통해 우리의 메소드의 효과를 이진 분류와 다중 분류 각각의 상황에서 증명하였다. 실험 결과는 우리의 메소드가 최근 준지도 학습 기법 연구들을 능가하는 것을 보여주며, 우리가 제안한 메소드가 정답이 없는 데이터를 학습에 효과적으로 활용하는 것을 보여준다. 더욱이, 표현 공간 상에서의 시각적 해석을 통해 우리의 메소드가 전반적으로 향상된 예측 결과와 클래스 별로 분리된 표현 공간을 구성한다는 것을 확인할 수 있었다.

주요어: 적대적 학습, 표현 분류자, 의료 영상 분할, 표현 학습, 준지도 학습
학번: 2021-20348