# SemSim: Revisiting Weak-to-Strong Consistency from a Semantic Similarity Perspective for Semi-supervised Medical Image Segmentation

Shiao Xie, Hongyi Wang, Ziwei Niu, Hao Sun, Shuyi Ouyang, Yen-Wei Chen, *Member, IEEE*, and Lanfen Lin, *Member, IEEE* 

Abstract - Semi-supervised learning (SSL) for medical image segmentation is a challenging yet highly practical task, which reduces reliance on large-scale labeled dataset by leveraging unlabeled samples. Among SSL techniques, the weak-to-strong consistency framework, popularized by FixMatch, has emerged as a state-of-the-art method in classification tasks. Notably, such a simple pipeline has also shown competitive performance in medical image segmentation. However, two key limitations still persist, impeding its efficient adaptation: (1) the neglect of contextual dependencies results in inconsistent predictions for similar semantic features, leading to incomplete object segmentation; (2) the lack of exploitation of semantic similarity between labeled and unlabeled data induces considerable class-distribution discrepancy. To address these limitations, we propose a novel semi-supervised framework based on FixMatch, named SemSim, powered by two appealing designs from semantic similarity perspective: (1) rectifying pixel-wise prediction by reasoning about the intra-image pair-wise affinity map, thus integrating contextual dependencies explicitly into the final prediction; (2) bridging labeled and unlabeled data via a feature querying mechanism for compact class representation learning, which fully considers cross-image anatomical similarities. As the reliable semantic similarity extraction depends on robust features, we further introduce an effective spatial-aware fusion module (SFM) to explore distinctive information from multiple scales. Extensive experiments show that SemSim yields consistent improvements over the state-of-the-art methods across three public segmentation benchmarks.

Index Terms—Semi-supervised learning, medical image segmentation, semantic similarity

#### I. INTRODUCTION

Medical image segmentation plays a critical role in various clinical scenarios, such as disease diagnosis and preoperative

Shiao Xie and Hongyi Wang contributed equally to this work.

Shiao Xie, Hongyi Wang, Ziwei Niu are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310063, China (e-mail: 22160144@zju.edu.cn, whongyi@zju.edu.cn, nzw@zju.edu.cn).

Hao Sun, Shuyi Ouyang are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310063, China (e-mail: sunhaoxx@zju.edu.cn, oysy@zju.edu.cn).

Yen-Wei Chen is with the College of Information Science and Engineering, Ritsumeikan University, Kusatsu 5250058, Japan, and also with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310063, China (e-mail: chen@is.ritsumei.ac.jp).

Lanfen Lin is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310063, China (e-mail: Ilf@zju.edu.cn).

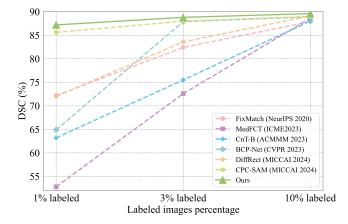


Fig. 1. Comparison of state-of-the-art methods with SemSim on the ACDC dataset under different labeled ratios.

assessment. In recent years, Convolutional Neural Networks (CNNs) have advanced the progress of this dense prediction task through the development of various segmentation models [1]–[4]. However, the laborious and time-consuming nature of manual annotation often results in a shortage of labeled data, impeding further progress in performance enhancement. Semi-supervised learning (SSL) has emerged as a notable solution, as it enables the utilization of large volume of unlabeled data, thus reducing the annotation burden significantly.

Several techniques are commonly used in SSL, including entropy minimization [5], pseudo-labeling [6]–[8], and consistency regularization [9]–[15]. In this realm, FixMatch [14] has garnered significant attention in classification tasks, which is grounded in weak-to-strong consistency regularization. Interestingly, the experimental results in Fig. 1 reveal that it also achieves comparable performance on medical image segmentation benchmarks, particularly in scenarios with extremely limited labeled data. Thus, we select this simple yet effective framework as our baseline. As shown in Fig. 2 (a), when adapting FixMatch [14] to segmentation tasks, the prediction  $p^w$  of the weak augmented view  $x^w$  acts as pseudo-label, enforcing pixel-wise consistency on the prediction  $p^s$  of the strong augmented view  $x^s$ . Intuitively, the success of this framework lies in generating high-quality  $p^w$ .

Unfortunately, there still exist two main issues that have been overlooked in this design: (i) Intra-image problem: According

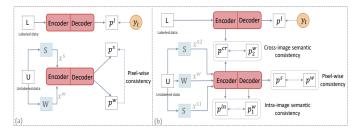


Fig. 2. Comparison of (a) FixMatch with (b) SemSim. S and W are strong and weak augmentations.  $y_l$  represents the label and  $p^l$  is the prediction of the labeled data.

to the label propagation algorithm [16], features with higher similarity should be assigned the same labels. FixMatch [14] only focuses on pixel-wise prediction and fails to uncover this implicit pair-wise correlations in feature space within an unlabeled image, leading to discontinuities and incomplete semantic features (shown in the left column of Fig. 3). (ii) Cross-image problem: Labeled and unlabeled data should follow the same category distribution. However, the limited volume of labeled data, along with its insufficient exploitation by conventional supervised training alone, often results in discrepancies between the class distributions learned from labeled and unlabeled data (shown in the right column of Fig. 3). In light of these problems, we conclude that *relying solely on existing consistency constraint is far from sufficient*.

To this end, we propose a novel SSL framework for medical image segmentation, named SemSim, from the perspective of semantic similarity. As illustrated in Fig. 2 (b), instead of feeding a single strong augmented view  $x^s$  into the model, we independently yield dual-stream perturbations  $(x^{s1}, x^{s2})$ from unlabeled data, accompanied by two additional consistency constraints: (i) Intra-image semantic consistency: The dependable contextual dependencies are critical for precise medical image segmentation. Motivated by this, we propose extracting the feature-level affinity map to refine the original pixel-wise prediction, and obtain more stable predictions,  $p^{in}$ for  $x^{s1}$  and  $p_1^w$  for  $x^w$ . Applying the constraint between  $p_1^w$ and  $p^{in}$  provides explicit guidance for deep semantic features, thus enhancing the feature continuity in Fig. 3. (ii) Crossimage semantic consistency: Medical imaging data exhibits anatomical structural similarities across different images, which creates a potential opportunity to establish a connection between labeled and unlabeled data. Therefore, we propose to directly intervene in the unlabeled data training flow with labeled data, and leverage the cross-image semantic feature similarity to generate predictions,  $p^{cr}$  for  $x^{s2}$  and  $p_2^w$  for  $x^w$ . The weak-tostrong consistency between  $p^{cr}$  and  $p_2^w$  in turn fully exploits reliable knowledge from labeled data thereby promoting a more compact intra-class distribution in Fig. 3.

Furthermore, it is evident that extracting powerful feature representations is vital for accurately calculating both intraand cross-image feature similarities. Previous studies have demonstrated that multi-scale features are able to capture distinctive information and address complex scale variations. Inspired by this, we develop a cross-scale feature fusion module that leverages the Transformer's [18], [19] ability to build long-range dependencies. Specifically, we take into account the spatial correspondences among patches at various

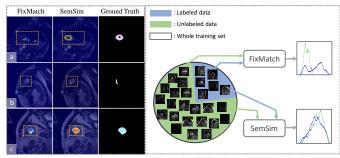


Fig. 3. Left: Visualization of class activation maps generated by Grad-CAM [17] for FixMatch and SemSim. (a), (b) and (c) represent features of class Myo, RV, and LV. Right: Kernel density estimations of voxels belonging to the Myo class in the ACDC dataset. FixMatch suffers from empirical distribution mismatch between labeled and unlabeled data, while SemSim effectively narrows the distribution gap.

scales through down-sampling operations, thus generating more representative features in a lightweight manner. Benefiting from the above designs, our framework achieves superior results over state-of-the-art (SOTA) semi-supervised medical image segmentation methods, as compared in Fig. 1. In summary, the main contributions are four-fold:

- (i) We analyze the semantic inconsistency issues ignored by FixMatch when extended to semi-supervised medical image segmentation. Our proposed framework, SemSim, aims to strengthen the reliability of semantic features as well as establish semantic consistencies from both intra- and crossimage perspectives.
- (ii) We introduce intra-image semantic consistency by imposing explicit constraints on contextual information and cross-image semantic consistency to establish cohesive class distributions.
- (iii) We design a lightweight spatial-aware fusion module to generate more powerful feature representations, enabling capture of more dependable correlations within the data.
- (iv) We conduct extensive experiments on three public medical image segmentation benchmarks, demonstrating that SemSim consistently outperforms other SSL methods.

## II. RELATED WORKS

#### A. Semi-supervised Learning

The main challenge in SSL lies in the design of effective and robust supervision signals for unlabeled data. Current approaches can be categorized into three major strategies: entropy minimization [5], pseudo-labeling [6]-[8], and consistency regularization [9]-[15]. Among these, the essence of consistency regularization is ensuring that a model's predictions for the same unlabeled sample remain consistent across different perturbations. FixMatch [14] stands out as a key framework based on consistency regularization. Building upon FixMatch [14], FlexMatch [20] integrates a curriculum learning strategy [21] that dynamically adjusts thresholds based on the learning state, leading to a significant reduction in training time. Additionally, FreeMatch [22] introduces a self-adaptive class fairness regularization penalty to promote diverse predictions in the early stages of training. Fast FixMatch [23] proposes utilizing fixed curriculum to regulate the unlabeled batch size, which reduces the computational load during training. These mentioned works, as variants of FixMatch [14], have

demonstrated considerable performance improvements in semisupervised classification tasks. Our proposed SemSim inherits from FixMatch [14] yet addresses its limitations by focusing on semantic consistency from both intra- and cross-image perspectives, improving its adaptability to segmentation tasks.

## B. Semi-supervised Medical Image Segmentation

Semi-supervised segmentation methods [11], [12], [15], [24]–[27] aim to leverage a substantial amount of unlabeled data to improve segmentation performance. One of the mainstream approaches is founded on the principle of consistency regularization. Mean-Teacher framework (MT) [9] and its variants [10] focus on allowing the predictions generated from either teacher or student model as close as possible. BCP-Net [28] suggests learning common semantics between the labeled and unlabeled data by a simple bidirectional copy-paste strategy. Another notable approach, CPS [6], extends pseudolabeling by using pseudo-labels generated by one perturbed segmentation network to supervise another. Recent researches highlight the exceptional performance of Transformers [18], [19], [29] in segmentation tasks. Inspired by this, CTCT [7] integrates CNN and Transformer architectures in a crossteaching manner, enabling the model to capture both local and global dependencies effectively. S4CVnet [8] introduces a dual-view co-training strategy along with consistency-aware supervision. MedFCT [27] achieves a more powerful backbone by efficiently exploring the fusion mechanism of Transformer and CNN in the Fourier domain. CnT-B [30] designs a bilevel uncertainty estimation strategy to ensure that the learning direction is more inclined towards the reliable side.

With the advent of large foundation models [31]–[33], DiffRect [34] capitalizes on the power of DDPM [33] to learn the latent structure of the semantic labels. CPC-SAM [35] leverages the prompting mechanism in SAM [31], which automatically generates prompts and supervisions across two decoder branches. To pursue simplicity and elegance, we use FixMatch [14] as the baseline and investigate how to fully unlock its potential in the medical image segmentation task.

#### III. METHODOLOGY

#### A. Prelimentaries

FixMatch [14] employs a weak-to-strong consistency regularization to leverage unlabeled data. The prediction  $p^w$  of an unlabeled image  $x_u$  with weak perturbation  $\mathcal{A}_w$  is used to constrain the prediction  $p^s$  of the same image with strong augmentation  $\mathcal{A}_s$ . A custom segmentation network denoted as F, such as UNet [2], is employed. The unsupervised loss  $\mathcal{L}_u$  can be formulated as:

$$p^{w} = F(\mathcal{A}_{w}(x_{u})), \quad p^{s} = F(\mathcal{A}_{s}(x_{u})),$$

$$\mathcal{L}_{u} = \frac{1}{|\mathcal{B}_{u}|} \sum \mathbb{1}(\max(p^{w}) \geq \tau) \odot \mathcal{L}_{dice}(p^{w}, p^{s}), \tag{1}$$

where  $\mathcal{B}_u$  is a batch of unlabeled data, and  $\tau$  is a pre-defined confidence threshold used to filter noisy pseudo labels. The supervised loss  $\mathcal{L}_s$  combines the cross-entropy loss  $\mathcal{L}_{ce}$  and

the dice loss  $\mathcal{L}_{dice}$  to minimize the difference between the prediction  $p^l$  and the ground truth  $y_l$ , it can be formulated as:

$$\mathcal{L}_s = \frac{1}{2} (\mathcal{L}_{ce}(p^l, y_l) + \mathcal{L}_{dice}(p^l, y_l)). \tag{2}$$

Then, the overall objective  $\mathcal{L}$  combining supervised loss  $\mathcal{L}_s$  and unsupervised loss  $\mathcal{L}_u$  can be computed as:

$$\mathcal{L} = \lambda_s \mathcal{L}_s + \lambda_u \mathcal{L}_u, \tag{3}$$

where  $\lambda_s$  and  $\lambda_u$  are weighting coefficients that balance the supervised loss and unsupervised loss, respectively.

## B. Intra-image Semantic Consistency

Contextual dependencies are crucial for assessing the model's training quality, where features extracted from a well-trained network should exhibit high similarity within the same class and being distinctly separable between different classes. However, in scenarios where labeled data is limited, relying on pixel-wise consistency constraint alone may be inadequate. It might fail to effectively guide a segmentation network in capturing reliable contextual dependencies from individual pixels. Thus, we opt to directly incorporate relationships extracted from the affinity map in feature-level to enhance the original pixel-wise prediction. The resulting predictions are referred as **intra-image semantic similarity-based predictions**.

As illustrated in Fig. 4 (a) and (b), for an unlabeled image, the strong augmented view  $x^{s1}$  is first sent into encoder  $h_{\theta}$ . The extracted multi-scale feature maps are then fed into the spatial-aware fusion module (described in section III-D) to generate the enhanced feature representation  $\tilde{f}_u^{s1} \in \mathbb{R}^{D \times H \times W}$ , where D is the channel dimension, H and W denote height and width. After that, we reshape it to the shape of  $D \times HW$  and compute the intra-image affinity map  $M \in \mathbb{R}^{HW \times HW}$  as:

$$M(k_1, k_2) = \operatorname{softmax} \left( \frac{\tilde{f}_u^{s1}(i_1, j_1)^{\top} \cdot \tilde{f}_u^{s1}(i_2, j_2)}{\left\| \tilde{f}_u^{s1}(i_1, j_1) \right\|_2 \left\| \tilde{f}_u^{s1}(i_2, j_2) \right\|_2} \right), \quad (4)$$

where (i., j.) is the coordinate of a pixel in the feature map, and (k., k.) is the coordinate in the affinity map. Note that  $k_1$  and  $(i_1, j_1)$  denote the position of the same pixel. M enables accurate delineation of the corresponding regions belonging to the same object.

To enhance the model's awareness of pairwise similarity, the refined prediction  $p^{in} \in \mathbb{R}^{C \times HW}$ , achieved by combining intra-class affinity reasoning result and the original prediction  $p^s$ , can be represented as:

$$p^{in} = I(p^s) + I(p^s) \cdot M, \tag{5}$$

where  $p^s \in \mathbb{R}^{C \times HW}$  is generated by segmentation decoder  $d_{\theta}$ .  $I(\cdot)$  is a bilinear interpolation for shape matching. For the weak augmented view  $x^w$ , the refined prediction  $p_1^w$  can also be computed with  $p^w$  in the same way. Finally, we perform consistency constraint  $\mathcal{L}_{intra}$  between  $p_1^w$  and  $p^{in}$ :

$$\mathcal{L}_{intra} = \frac{1}{|\mathcal{B}_u|} \sum \mathbb{1}(\max(p_1^w) \ge \tau) \odot \mathcal{L}_{dice}(p_1^w, p^{in}). \quad (6)$$

In this way, we explicitly spread the contextual dependencies into model logits output, and optimize them through a weakto-strong consistency constraint.

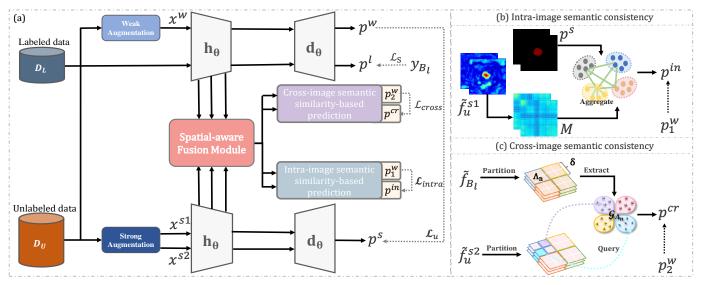


Fig. 4. (a) Overview of SemSim framework. (b) The predictions  $p^{in}$ ,  $p_1^w$  are based on intra-image semantic similarity. (c) The predictions  $p^{cr}$ ,  $p_2^w$  are based on cross-image semantic similarity.

# C. Cross-image Semantic Consistency

The insufficient utilization of limited labeled data makes it challenging for the model to achieve consistent class representations. Hence, we propose explicitly guiding the learning of category-wise semantic features from both labeled and unlabeled data. Specifically, we suggest generating predictions for the unlabeled data by leveraging the trustworthy class distribution derived from the labeled data, via a dynamic feature querying mechanism. This process accelerates the transfer of dependable information from labeled data to unlabeled data, thereby narrowing the distribution gap between them. The obtained predictions are defined as **cross-image semantic similarity-based predictions**. Each step will be explained in detail in the following sections.

1) Class Statistics of Labeled Batch  $\mathcal{B}_1$ : As shown in Fig. 4 (c), for labeled data, once the enhanced feature map  $\tilde{f}_{\mathcal{B}_l} \in \mathbb{R}^{|\mathcal{B}_l| \times D \times H \times W}$  is obtained, a typical method is to compute class prototypes directly based on masks, using this global information to represent the overall class distribution. However, due to the spatially heterogeneous characteristics of background, which involves various anatomical structures, this global operation will lead to an unreasonable averaging of structural information. Hence, given that medical images generally exhibit similar spatial layouts across different slices, we propose to partition the feature map into multiple subregions and calculate local class prototypes instead. As shown in Fig. 4 (c), we divide  $\tilde{f}_{\mathcal{B}_t}$  into a set of sub-regions  $\Omega = \{\Lambda_n\}_{n=1}^N$ with the region size of  $\delta \times \delta$ , N is the number of sub-regions. Within each  $\Lambda_n$ , the prototype  $\mathcal{G}_{\Lambda_n}^c$  for each class c is computed by averaging all features of the same class, which can be formulated as:

$$\mathcal{G}_{\Lambda_n}^c = \frac{\sum_{i \in \Lambda_n} \mathbb{1}(y_{\mathcal{B}_l}(i) = c) \odot \tilde{f}_{\mathcal{B}_l}(i)}{\sum_{i \in \Lambda} \mathbb{1}(y_{\mathcal{B}_l}(i) = c)},\tag{7}$$

where  $y_{\mathcal{B}_l}$  is the class label of the entire  $\mathcal{B}_l$  and  $\mathcal{G}_{\Lambda_n} \in \mathbb{R}^{|\mathcal{B}_l| \times C \times D}$ . In that way, the class prototypes obtained tends to be more accurate and reliable than global ones.

2) Dynamically Query Labeled Batch  $\mathcal{B}_1$ : As shown in Fig. 4 (c), for an unlabeled image, after we obtain its enhanced feature map  $\tilde{f}_u^{s2} \in \mathbb{R}^{D \times H \times W}$ , we also perform the same region partition as labeled feature maps. Then, we may employ the cosine similarity between the query feature in  $\tilde{f}_u^{s2}$  and the corresponding matched sub-region prototype set  $\mathcal{G}_{\Lambda_n}$ :

$$m_{\Lambda_{-}}(i,j) = \cos(\tilde{f}_{s_{-}}^{s_{2}}(i), \mathcal{G}_{\Lambda_{-}}(j)), \tag{8}$$

where  $i \in \Lambda_n$  and  $j \in |\mathcal{B}_l|$ . Accordingly, we will obtain similarity score vector at the i-th position  $m_{\Lambda_n}(i) \in \mathbb{R}^{|\mathcal{B}_l| \times C}$ , which reflects the affinity between this feature vector and the prototypes within all batch corresponding to its sub-region. Similarly, we can apply the same operation to other sub-regions and calculate the overall similarity map  $m \in \mathbb{R}^{|\mathcal{B}_l| \times C \times H \times W}$ . The final cross-image semantic similarity based prediction  $p^{cr} \in \mathbb{R}^{C \times H \times W}$  can be defined as:

$$p^{cr}(c) = \frac{\sum_{j=1}^{|\mathcal{B}_l|} e^{m(j,c)}}{\sum_{c} \sum_{j=1}^{|\mathcal{B}_l|} e^{m(j,c)}}.$$
 (9)

*3) Uncertainty Estimation of* p<sup>cr</sup>: In the initial phase of training, feature learning can exhibit considerable uncertainty. To effectively assess the stability of the training process and the consistency of predictions, we evaluate the uncertainty of the obtained cross-image semantic similarity-based predictions. Specially, we first compute the average prediction by querying a batch of labeled data:

$$\bar{m} = \frac{1}{|\mathcal{B}_l|} \sum_{j=1}^{|\mathcal{B}_l|} m(j). \tag{10}$$

The uncertainty of the j-th prediction by inferring the j-th labeled image can be calculated by the KL-divergence (KL):

$$U(j) = \text{KL}[m(j)||\bar{m}] = \sum_{c=0}^{C-1} m(j,c) \log \frac{m(j,c)}{\bar{m}(c)}.$$
 (11)

Such pixel-level uncertainty reflects the approximate variance that assesses the difference between the batch-wise and the averaged one, where a larger value indicates lower similarity.

The average uncertainty  $\bar{U} \in \mathbb{R}^{H \times W}$  across batch of labeled images for the prediction  $p^{cr}$  can be defined as:

$$\bar{U} = \frac{1}{|\mathcal{B}_l|} \sum_{j=1}^{|\mathcal{B}_l|} e^{-r \times U(j)}, \tag{12}$$

where r is a hyperparameter. The corresponding weak-tostrong consistency loss will be enforced on this prediction by incorporating the uncertainty computed before:

$$\mathcal{L}_{cross} = \frac{1}{|\mathcal{B}_u|} \sum \mathbb{1}(\max(p_2^w) \ge \tau) \odot \mathcal{L}_{ce}(p_2^w, p^{cr}) \odot \bar{U},$$
(13)

where  $p_2^w$  is also obtained by calculating the cross-image semantic similarity. It should be noted that when the differences between labeled and unlabeled batches are significant, e.g. the local prototype of a specific class is absent in a certain subregion, we will prioritize using the class prototype of the entire image as a substitute.

#### D. Spatial-aware Fusion Module

Multi-scale cues have been shown to be highly effective in addressing complex scale variations, allowing for the capture of both global and local contextual information of the target. The key challenge lies in how to efficiently integrate this information to construct a robust feature representation, which is essential for the reliability of our proposed semantic similarity-based predictors. To tackle this, we develop a spatial-aware transformer-based block that facilitates interaction across different scales in an efficient manner. Fig. 5 illustrates the specific fusion process from the i-th to the (i+2)-th scale.

Rather than computing dependencies within a lengthy sequence directly concatenated from tokens at multiple scales (obtained by flattening the entire feature map of each scale), we focus on the long-range cross-scale dependency of the object itself and the nearby objects. First, we harness the spatial relationships among patches in different scales, called **Patch Matching**. As shown in Fig. 5, we denote  $T_j^i$  as the j-th patch at the i-th scale, and the corresponding down-sampled patches are  $T_j^{(i+1)}$  and  $T_j^{(i+2)}$  sharing the bounding box with the same color (e.g. yellow). We first regularize feature maps into the same channel dimension, then concatenate the corresponding inter-scale patches, which can be formulated as:

$$\left[ \text{flatten}\left(T_j^i\right), \text{flatten}\left(T_j^{(i+1)}\right), \text{flatten}\left(T_j^{(i+2)}\right) \right] \to T_j^{\text{cat}},$$
(14)

where  $flatten(\cdot)$  rearranges the patch size into 1D sequence and  $[\cdot]$  represents the concatenation operation. In this way, we focus on the most related patches, and thus maintain the spatial correspondence as well as reduce the redundancy. The second step *Scale Interacting* that aims to capture dependencies among patches can obtained as follows:

$$\hat{T}_{j}^{cat} = \text{MSA}\left(\text{LN}\left(T_{j}^{cat}\right)\right) + T_{j}^{cat}, 
\tilde{T}_{j}^{cat} = \text{MLP}\left(\text{LN}\left(\hat{T}_{j}^{cat}\right)\right) + \hat{T}_{j}^{cat},$$
(15)

where  $LN(\cdot)$  denotes the Layer Normalization [36],  $MSA(\cdot)$  represents Multi-head Self-Attention operation and  $MLP(\cdot)$  denotes a two-layer linear projection. After that, we will utilize

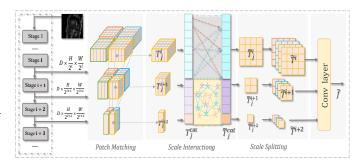


Fig. 5. Overview of spatial-aware fusion module. H, W represent the height and width of the feature map.

*Scale Splitting* to convert the enhanced sequence back to patches based on the concatenation order:

Reshape 
$$\left( \operatorname{Split} \left( \tilde{T}_{j}^{\operatorname{cat}} \right) \right) \to \tilde{T}_{j}^{i}, \tilde{T}_{j}^{(i+1)}, \tilde{T}_{j}^{(i+2)},$$
 (16)

where  $\mathrm{Split}(\cdot)$  is the inverse operation of previous concatenation operation. Finally, we perform interpolation  $(I(\cdot))$  on the enhanced feature maps to uniform the spatial dimension, and then employ convolution layers to fuse these interpolated feature maps and obtain the final feature representation  $\tilde{f}$ . This process is described as follows:

$$\tilde{f} = \text{RELU}(\text{BN}(\text{Conv}([I(\tilde{T}^i), \tilde{T}^{(i+1)}, I(\tilde{T}^{(i+2)})]))), \quad (17)$$

where  $BN(\cdot)$  represents BatchNormalization layer [37] and  $RELU(\cdot)$  [38] is an activation function.

Complexity Analysis. If we directly concatenate the multiscale feature maps and apply an original transformer block, the computational complexity is  $O(\frac{H^2W^2D}{16^i})$ , where the sequence length is  $(\frac{HW}{2^{2i}} + \frac{HW}{2^{2i+2}} + \frac{HW}{2^{2i+4}})$ . Our proposed spatial-aware fusion module first splits the feature map at each scale into  $S \times S$  windows, with each window having a size of  $\frac{H}{2^i \times S} \times \frac{W}{2^{ii} \times S}$  in the i-th stage. By leveraging cross-scale spatial-aware correspondences, we only need to perform self-attention  $S \times S$  times, with a sequence length given by  $\frac{1}{S^2} \left( \frac{HW}{2^{2i}} + \frac{HW}{2^{2i+2}} + \frac{HW}{2^{2i+4}} \right)$  for each operation. Thus, the complexity of each operation is approximately  $O\left(\frac{H^2W^2D}{16^iS^2}\right)$ , and the total computational load becomes  $O\left(\frac{H^2W^2D}{16^iS^2}\right)$ .

#### E. Training and Inference

1) Loss Function: In summary, based on the original unsupervised loss (Eq. 1) in FixMatch [14], our SemSim further incorporates two additional consistency losses from semantic similarity perspective:  $\mathcal{L}_{intra}$  (Eq. 6) and  $\mathcal{L}_{cross}$  (Eq. 13). Therefore, the unsupervised loss can be rephrased as:

$$\tilde{\mathcal{L}}_{u} = \lambda \mathcal{L}_{u} + \lambda_{intra} \mathcal{L}_{intra} + \lambda_{cross} \mathcal{L}_{cross}$$
 (18)

where the loss weights  $\lambda$ ,  $\lambda_{intra}$ ,  $\lambda_{cross}$  are set to 0.5, 0.25, 0.25, respectively. Consequently, the overall objective can be described in the following, where the supervised loss  $\mathcal{L}_s$  stays consistent with FixMatch [14]:

$$\mathcal{L} = \frac{1}{2}(\mathcal{L}_s + \tilde{\mathcal{L}}_u). \tag{19}$$

#### **Algorithm 1:** Pseudocode of SemSim.

2) The Algorithm of SemSim: The pseudocode of our SemSim framework is presented in Algorithm III-D. This framework takes an equal number of labeled and unlabeled images as input, and aims to obtain the fully-trained F for inference. During inference process, assuming the test input image is  $x_t$ , we just need to pass it through the encoder  $h_\theta$  and decoder  $d_\theta$  of F to obtain the final prediction.

#### IV. EXPERIMENTS

#### A. Datasets and Evaluation Metrics

In our experiments, we use three commonly used datasets to verify the effectiveness of our method, including ACDC dataset [39], ISIC dataset [40] and PROMISE12 dataset [41].

- (i) ACDC dataset is a cardiac MRI dataset that contains 200 annotated short-axis cardiac MRI images from 100 patients with three organs. Following [7], we split the dataset into 70 patients for training, 10 patients for validating and 20 patients for testing, respectively. For semi-supervised training, we evaluate different methods with 1% (1 case), 5% (3 case) and 10% (7 case) labeled separately.
- (ii) ISIC dataset is a skin lesion segmentation dataset with 2 classes. We use 1838 images for training and the rest 756 images for validation. Under a semi-supervised setting, 3% (55 images), 10% (181 images) of training data are provided with labels, while the rest of training images are unlabeled.
- (iii) PROMISE12 dataset is available for the MICCAI 2012 prostate segmentation challenge. T2-weighted MRIs of 50 patients with various conditions are acquired at different locations. The dataset is divided into 35 training cases, 5 validation cases, and 10 test cases. All 3D scans are converted into 2D slices. For the semi-supervised setting, 3% and 7% are regarded as labeled, while the rest remain unlabeled.

Three commonly used metrics are employed to evaluate the segmentation results: (1) Dice Coefficient (DSC), (2) 95% Hausdorff Distance (95HD), and (3) Average Symmetric Surface Distance (ASSD). The DSC measures the overlap between the prediction and the ground truth, while the 95HD calculates the distances between the surfaces of the prediction and the ground truth. Following previous approaches [42], DSC

and 95HD are used to evaluate performance on the ACDC and ISIC datasets, whereas DSC and ASSD are utilized as evaluation metrics for the PROMISE12 dataset.

# B. Implementation Details

Following [7], UNet [2] is adopted as the backbone. We resize all the slices to  $224 \times 224$ . Our augmentation strategy is consistent with [7], where color transformation and CutMix [43] are set as strong augmentations, and random flipping is utilized as weak augmentation. During training, the batch size is set to 16, consisting of an equal number of unlabeled and labeled cases. We optimize our network with the SGD optimizer [44], where the weight decay is set to  $1 \times 10^{-4}$  and the momentum is set to 0.9. All experiments are trained for 300 epochs. Moreover, the learning rate is initially set to 0.01. To broaden the perturbation space, we further introduce feature level augmentation by adopting a channel dropout of 50% probability. The channel dimension D is set to 128 in our setting. r and  $\tau$  are set to 1000 and 0.95.

#### C. Comparison with State-of-the-Art

1) Quantitative Comparison: We compare our proposed Sem-Sim with 20 state-of-the-art semi-supervised learning methods, including (1) CNN-based methods: MT [9], UA-MT [10], EM [5], DCT [24], CCT [25], CPS [6], ICT [11], DAN [26], URPC [13], SSNet [15], ICT-Med [12], FixMatch [14], SCPNet [42], BCP-Net [28]; (2) hybrid methods (CNN and Transformer): CTCT [7], S4CVnet [8], MedFCT [27], CnT-B [30]; (3) generative method: DiffRect [34]; and (4) SAM-based method: CPC-SAM [35]. Following [7], we conduct all experiments under the same settings on the public ACDC, ISIC, and PROMISE12 datasets. "Only sup" in Tables I, II, and III refers to training with labeled data only.

**Results on ACDC dataset:** The quantitative comparison results on the ACDC dataset are shown in Table I. Specifically, with 5% and 10% labeled data, SemSim demonstrates remarkable improvements compared to the previous state-of-the-art method CPC-SAM [35] (DSC: +0.8%, +0.6%). Notably, even with only 1% labeled data, SemSim still substantially

TABLE I

COMPARISON OF SEMSIM WITH OTHER SSL METHODS ON ACDC DATASET UNDER DIFFERENT RATIOS OF LABELED DATA. — REPRESENTS THAT THE VALUE IS TOO UNSTABLE AND LOW TO WRITE, WHICH REMAINS BELOW 0.3 AFTER RETRAINING MULTIPLE TIMES.

Method	1% labeled		5% labeled		10% labeled		Params (M)
Method	DSC ↑	95HD (mm) ↓	DSC ↑	95HD (mm) ↓	DSC ↑	95HD (mm) ↓	Faranis (WI)
Only sup	0.390	54.7	0.560	39.8	0.797	9.8	2.60
MT [9]	-	-	0.566	34.5	0.810	14.4	1.81
EM [5]	-	-	0.602	24.1	0.791	14.5	1.81
UA-MT [10]	-	-	0.610	25.8	0.815	14.4	1.81
DCT [24]	-	-	0.582	26.4	0.804	13.8	1.81
CCT [25]	-	-	0.586	27.9	0.816	13.1	3.71
CPS [6]	-	-	0.603	25.5	0.833	11.0	3.62
ICT [11]	-	-	0.581	22.8	0.811	11.4	1.81
DAN [26]	-	-	0.528	32.6	0.795	14.6	1.81
URPC [13]	-	-	0.567	31.4	0.829	10.6	1.83
CTCT [7]	-	-	0.704	12.4	0.864	8.6	29.11
SSNet [15]	-	-	0.705	17.4	0.853	10.6	1.81
ICT-Med [12]	-	-	0.563	22.6	0.837	13.1	1.81
S4CVNet [8]	0.534	37.2	0.731	5.1	0.873	3.9	55.31
FixMatch [14]	0.722	22.8	0.824	4.5	0.879	3.1	1.81
MedFCT [27]	0.528	24.1	0.726	10.5	0.886	4.3	31.18
CnT-B [30]	0.632	19.3	0.755	10.6	0.880	5.5	29.11
BCP-Net [28]	0.649	18.0	0.879	2.1	0.889	4.0	1.81
DiffRect [34]	0.720	5.8	0.836	7.6	0.891	3.9	19.37
CPC-SAM [35]	0.856	9.2	0.880	5.8	0.890	3.1	93.75
SemSim <sup>-</sup>	0.861	4.3	0.882	2.4	0.893	2.7	1.81
SemSim	0.872	1.8	0.888	1.9	0.896	2.3	2.60

TABLE II

COMPARISON OF SEMSIM WITH OTHER SSL METHODS ON ISIC DATASET UNDER DIFFERENT RATIOS OF LABELED DATA.

Method	39	% labeled	10% labeled		
Method	DSC ↑	95HD (mm) ↓	DSC ↑	95HD (mm) ↓	
Only sup	0.663	28.4	0.691	26.3	
MT [9]	0.728	37.4	0.734	34.0	
EM [5]	0.723	36.3	0.727	39.3	
UA-MT [10]	0.730	38.6	0.734	33.2	
DCT [24]	0.729	40.6	0.760	35.7	
CCT [25]	0.677	42.2	0.723	31.7	
CPS [6]	0.686	44.4	0.743	35.7	
ICT [11]	0.732	37.2	0.753	34.6	
DAN [26]	0.695	39.5	0.724	30.4	
URPC [13]	0.703	39.3	0.758	32.8	
CTCT [7]	0.713	43.2	0.760	37.3	
SSNet [15]	0.728	40.8	0.758	32.8	
ICT-Med [12]	0.714	39.2	0.749	33.1	
S4CVNet [8]	0.752	37.2	0.774	31.8	
MedFCT [27]	0.742	39.0	0.780	29.6	
FixMatch [14]	0.756	36.7	0.783	15.5	
CnT-B [30]	0.762	29.6	0.792	25.0	
BCP-Net [28]	0.769	24.5	0.794	18.7	
SemSim	0.774	18.9	0.802	13.3	

outperforms existing approaches in both the regional measure DSC (+1.6%) and the boundary-aware measure 95HD (-7.4mm). This highlights that our method offers a significant advantage in scenarios with minimal labeled data, where most of the other methods perform poorly (with DSC below 0.3). Besides, our SemSim contains only 2.60 million parameters, substantially fewer than several high-performing methods such as DiffRect [34] and CPC-SAM [35].

We further eliminate the interference of spatial-aware fusion

TABLE III

COMPARISON OF OUR SEMSIM WITH OTHER SSL METHODS ON 3% AND
7% LABELED DATA OF PROMISE12 DATASET.

Method	39	% labeled	7% labeled		
Method	DSC ↑ ASSD (mm) ↓		DSC ↑	ASSD (mm)↓	
Only sup	0.552	12.8	0.587	8.4	
MT [9]	0.405	3.1	0.714	7.6	
UA-MT [10]	0.552	8.6	0.657	2.4	
CCT [25]	0.467	2.1	0.714	16.7	
URPC [13]	0.619	2.9	0.632	4.3	
SSNet [15]	0.574	6.3	0.623	4.4	
SCP-Net [42]	0.551	19.5	0.771	3.5	
S4CVnet [8]	0.628	1.9	0.672	3.1	
FixMatch [14]	0.463	4.2	0.716	6.7	
CnT-B [30]	0.685	5.4	0.763	2.7	
MedFCT [27]	0.725	2.2	0.760	1.8	
BCP-Net [28]	0.706	8.8	0.772	1.4	
SemSim	0.758	1.6	0.784	1.3	

module (SFM) and replace it with a simpler convolution (SemSim<sup>-</sup>). It can be observed that even without this module, our model equipped with intra- and cross-image semantic consistencies already outperforms existing methods in all partitions. Due to the lightweight design of SFM, which consists of only 0.79 million parameters, SemSim achieves significant improvement, particularly with 1% labeled data.

**Results on ISIC dataset:** Table II presents the quantitative results of different methods on the ISIC dataset. Our framework, compared to training with only labeled data (Only sup), effectively leverages unlabeled data to enhance segmentation accuracy. Consistent with the results in Table I, SemSim outperforms the previous best method, BCP-Net [28], with

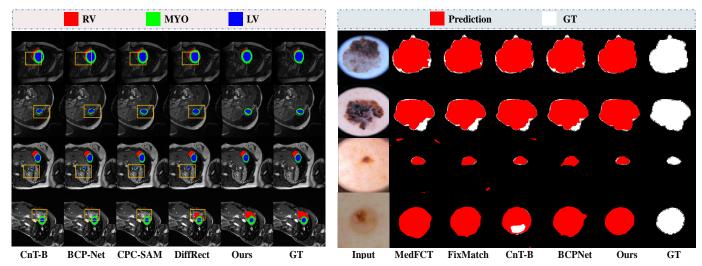


Fig. 6. Comparison of segmentation results on the ACDC (left) and ISIC (right) datasets with 10% labeled data.

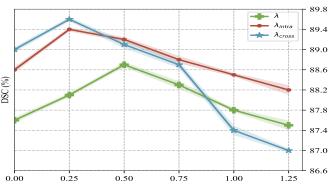


Fig. 7. Performance of SemSim w.r.t.  $\lambda, \lambda_{intra}$  and  $\lambda_{cross}$  on the ACDC dataset.

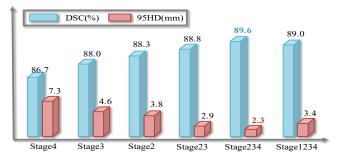


Fig. 8. Ablation of spatial-aware fusion module on different stages on the ACDC dataset with 10% labeled.

improvements in DSC (+0.5% and +0.8%) and reductions in 95HD (-5.6mm and -5.4mm) under the 3% and 10% labeled data settings, respectively, establishing a new state-of-the-art. These results on both datasets substantiate the robustness and effectiveness of our method.

Results on PROMISE12 dataset: Table III presents the performance comparison between SemSim and other SOTA methods on the PROMISE12 dataset, using labeled data percentages of 3% and 7%. Compared with training with only labeled data, SemSim can achieve +20.6% and 19.7% improvements with 3% and 7% labeled. Although recent approaches such as MedFCT [27] and BCP-Net [28] have shown promising results, they fall short in effectively delineating clear boundaries and establishing distinct category relationships. Our SemSim, which integrates newly proposed intra- and cross-image semantic

TABLE IV
ABLATION STUDY OF DIFFERENT CONSISTENCY CONSTRAINTS ON 1%
LABELED DATA OF ACDC DATASET.

#	$\mathcal{L}_{intra}$	$\mathcal{L}_{cross}$	$\mathcal{L}_u$	DSC ↑	95HD (mm) ↓
1	×	×	<b>√</b>	0.835	7.6
2	✓	×	×	0.844	5.0
3	×	✓	×	0.824	5.5
4	✓	✓	×	0.861	4.3
5	✓	×	✓	0.868	2.9
6	×	✓	✓	0.862	4.1
7	✓	✓	$\checkmark$	0.872	1.8

TABLE V EFFECT OF THE NUMBER OF SUB-REGIONS  $m{N}$  ON ACDC, ISIC AND PROMISE12 DATASETS.

N	AC	CDC (10%)	IS	SIC(10%)	PROMISE12 (7%)	
11	DSC ↑	95HD (mm) ↓	DSC ↑	95HD (mm) ↓	DSC ↑	ASSD (mm) ↓
1	0.888	3.6	0.775	18.1	0.770	3.4
2	0.891	2.9	0.795	16.5	0.778	2.1
4	0.896	2.3	0.802	13.3	0.784	1.3
8	0.893	2.4	0.794	15.3	0.780	2.9
14	0.887	4.3	0.789	17.1	0.775	4.5

consistency constraints with a powerful feature fusion module, achieving the highest performance 75.8% and 78.4% under 3% and 7% labeled settings on this dataset.

2) Visual Comparisons: Fig. 6 illustrates the qualitative results of different methods on the ACDC dataset [39] with 10% labeled and ISIC dataset [40] with 10% labeled, including FixMatch [14], MedFCT [27], CnT-B [30], BCPNet [28], CPC-SAM [35], DiffRect [34], ours and GroundTruth. Benefiting from the intra- and cross-image semantic consistencies design, SemSim is capable of generating satisfactory segmentation results closer to the ground truth with only 10% labeled, ensuring both the integrity and accuracy of the segmented targets. The phenomenon can be explained by its ability to enhance feature continuity and strengthen the distinctions between categories.

TABLE VI
ABLATION STUDY OF STRONG PERTURBED STREAMS ON ACDC AND
ISIC DATASETS WITH 10% LABELED DATA.

$p^{in}$	$p^{cr}$	$p^s$	ACDC	(10% labeled)	ISIC (10% labeled)		
Р			DSC ↑	95HD (mm) ↓	DSC ↑	95HD (mm) ↓	
S1	S1	S1	0.890	2.0	0.790	20.0	
<b>S</b> 1	S1	S2	0.884	2.7	0.786	19.8	
<b>S</b> 1	S2	S1	0.896	2.3	0.802	13.3	
<b>S</b> 1	S2	S2	0.877	1.9	0.779	15.8	
<b>S</b> 1	S2	S3	0.887	1.5	0.755	31.1	
×	×	S1/S2	0.868	2.1	0.780	19.9	
×	×	S1/S2/S3	0.889	1.4	0.795	16.5	

# D. Hyper-Parameters

1) Impact of Loss Function Weights:  $\lambda$ ,  $\lambda_{intra}$  and  $\lambda_{cross}$  are three coefficients that balance the overall unsupervised loss. We explore the sensitivity of the remaining weight by fixing two of them. As shown in Fig. 7, we first fix  $\lambda_{intra} = 0.25$  and  $\lambda_{cross} = 0.25$ , it is evident that the accuracy peaks at  $\lambda = 0.5$ , highlighting that the conventional pixel-wise consistency plays a more significant role in model performance. Next, with  $\lambda_{cross} = 0.25$  and  $\lambda = 0.5$ , the optimal performance of 89.4% is achieved at  $\lambda_{intra} = 0.25$ . Finally, we examine the appropriate value for  $\lambda_{cross}$  while fixing  $\lambda = 0.5$  and  $\lambda_{intra} = 0.25$ . As the weight  $\lambda_{cross}$  increases, we observe a sharp decline in accuracy primarily due to the overfitting of unlabeled data to the limited labeled data. Thus, the optimal combination of loss weights  $\lambda$ ,  $\lambda_{intra}$ ,  $\lambda_{cross}$  is 0.5, 0.25, 0.25.

2) Spatial-aware Fusion Module on Different Stages: Additionally, we examine the impact of the spatial-aware fusion module on performance across various scales, ranging from stage 1 (size of 224 × 224) to stage 4 (size of 28 × 28). As illustrated in Fig. 8, leveraging features from shallower stages leads to an improvement from 86.7% to 88.3%, as it encompasses both local details and adequate semantic information. The method achieves its highest performance (DSC: 89.6%, 95HD: 2.3mm) when applying multi-scale fusion across consecutive stages (stage 2, 3 and 4). However, incorporating fusion at stage 1 can compromise intrinsic characteristics and increase computational overhead. Therefore, we implement multi-scale fusion at stages 2, 3 and 4, resulting in optimal performance while maintaining a reasonable computational complexity.

# E. Ablation Study

1) Effect of Different Consistency Losses: We conduct experiments to explore the effectiveness of different consistency constraints. In Table IV, we begin by assessing the influence of intra-image semantic consistency constraint  $\mathcal{L}_{intra}$  (#2) and cross-image semantic consistency constraint  $\mathcal{L}_{cross}$  (#3). It indicates that with  $\mathcal{L}_{intra}$  only (#2), our performance has already surpassed that of  $\mathcal{L}_u$  (#1). Then, relying solely on the correlation with limited labeled data (#3), our method can still achieve competitive performance (DSC: 0.824, 95HD: 5.5mm), which validates the effectiveness of  $\mathcal{L}_{cross}$ . By combining these two types of consistency constraints (#4), our method

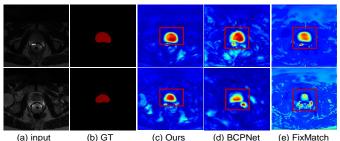


Fig. 9. Feature visualizations of different methods on PROMISE12 dataset.

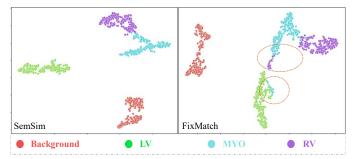


Fig. 10. T-SNE visualization of deep feature representations extracted from SemSim and FixMatch on ACDC dataset.

improves by +1.7% and +3.7% compared to using (#2 and #3) separately, demonstrating the complementarity of  $\mathcal{L}intra$  and  $\mathcal{L}_{cross}$ . Moreover, we further individually integrate  $\mathcal{L}_{cross}$  and  $\mathcal{L}_{intra}$  with  $\mathcal{L}_u$  (#5 and #6), performances have both been improved compared with using only  $\mathcal{L}_u$  (#1). Finally, SemSim achieves the best performance by combining these three consistency losses.

2) Effect of the Number of Sub-regions: We explore the impact of the number of sun-regions N on these three datasets, it can be observed from the Table V that when the number is set to 4, the best performances are achieved across all three datasets. As the number increases, there is an escalation in computational demand, while the accuracy experiences a corresponding decline, which could be attributed to the small size of the grid regions, making it challenging for unlabeled images to capturing accurate and representative class-specific features from such limited areas.

3) Ablation of Strong Perturbation Streams: We also evaluate the impact of providing different strong augmented views as input to the three types of predictions. As shown in Table VI, S1, S2, and S3 represent views subjected to varying degrees of strong data augmentation. The results indicate that using the same view for both  $p^s$  and  $p^{in}$  yields the highest accuracy on the ACDC and ISIC datasets. Besides, when predictions are made solely through the decoder, the triple-view setup (S1/S2/S3) outperforms the dual-view setup (S1/S2) due to the introduction of diverse perturbations. However, they are still inferior to our proposed comprehensive predictions integrating both intra- and cross-semantic similarity-based predictions, which fully leverage image-level perturbations.

4) Ablation of Consistency Constraint Framework: As illustrated in Fig. 11, we can obtain three types of predictions: through decoder  $(p^w, p^s)$ , intra-image semantic similarity-based predictions  $(p_1^w, p^{in})$ , cross-image semantic similarity-based predictions  $(p_2^w, p^{cr})$ . Then, we analyze the combinations

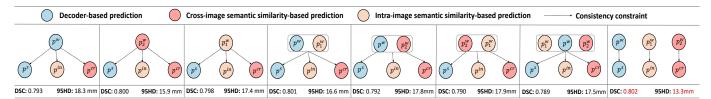


Fig. 11. Ablation of consistency constraint framework on the ISIC dataset with 10% labeled.

among these predictions to achieve three types of weak-to-strong consistencies  $\mathcal{L}_{intra}$ ,  $\mathcal{L}_{cross}$  and  $\mathcal{L}_u$ . Our results indicate that averaging  $p_1^w$  and  $p^w$  as the ensemble prediction for  $x^w$  achieves superior performance, with DSC: 0.801 and 95HD: 16.6mm. Notably, for each prediction type, the corresponding weak view prediction acts as a pseudo label to guide the strongly augmented predictions, yielding the highest accuracy of 0.802 on the ISIC dataset.

#### F. Interpretation of our SemSim

We visualize the class activation maps of various methods in Fig. 9. Compared to BCPNet [28] and FixMatch [14], SemSim inherits the advantages of intra-image semantic consistency that considers the relationships between pixels, reducing pixel-level misclassification and resulting in more complete object segmentation. As illustrated in Fig. 10, the pixel embeddings learned by SemSim become more compact and well-separated. It demonstrates that the designed crossimage semantic consistency effectively extracts distinctive features from the labeled data, even with a limited amount of data available.

#### V. CONCLUSION

In this paper, we propose a novel semi-supervised medical image segmentation framework, termed SemSim, which addresses the intra- and cross-image semantic inconsistency challenges faced by FixMatch. We thoroughly investigate two key consistency mechanisms based on semantic similarity: one emphasizes contextual dependencies between internal features within an image to refine final predictions, thereby achieving more continuous segmentation. The other extracts semantic relationships between labeled and unlabeled data, enabling the model to learn more accurate and consistent class distributions even with limited labeled data. Further, an efficient spatial-aware fusion module is introduced to assist the above consistencies by generating powerful feature representations. SemSim was extensively evaluated on three public datasets and consistently outperformed other SSL approaches.

#### REFERENCES

- J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on* computer vision and pattern recognition, 2015, pp. 3431–3440.
- [2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [3] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

- [4] M. Firdaus-Nawi, O. Noraini, M. Sabri, A. Siti-Zahrah, M. Zamri-Saad, and H. Latifah, "Deeplabv3+ \_encoder-decoder with atrous separable convolution for semantic image segmentation," *Pertanika J. Trop. Agric. Sci.*, vol. 34, no. 1, pp. 137–143, 2011.
- [5] T. Vu, H. Jain, and M. Bucher, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.
- [6] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2613–2622.
- [7] X. Luo, M. Hu, T.Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between cnn and transformer," arXiv preprint arXiv:2112.04894, 2021.
- [8] Z. Wang, T. Li, J.-Q. Zheng, and B. Huang, "When cnn meet with vit: Towards semi-supervised learning for multi-class medical image semantic segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 424–441.
- [9] A.Tarvainen and H.Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," Advances in Neural Information Processing Systems, vol. 30, 2017.
- [10] L. Yu, S. Wang, S. Li, C. Fu, and P. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 605–613.
- [11] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," arXiv preprint arXiv:1903.03825, 2019.
- [12] H. Basak, R. Bhattacharya, R. Hussain, and A. Chatterjee, "An embarrassingly simple consistency regularization method for semi-supervised medical image segmentation," arXiv preprint arXiv:2202.00677, 2022.
- [13] X. Luo, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, and S. Zhang, "Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 2021, pp. 318–329.
- [14] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semisupervised learning with consistency and confidence," *Advances in neural* information processing systems, vol. 33, pp. 596–608, 2020.
- [15] Y. Wu, Z. Wu, Q. Wu, Z. Ge, and J. Cai, "Exploring smoothness and class-separation for semi-supervised medical image segmentation," arXiv preprint arXiv:2203.01324, 2022.
- [16] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *Proceedings of the IEEE/CVF conference* on computer vision and pattern recognition, 2019, pp. 5070–5079.
- [17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international* conference on computer vision, 2017, pp. 618–626.
- [18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
- [19] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.
- [20] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing* Systems, vol. 34, pp. 18408–18419, 2021.
- [21] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference* on *Machine Learning*, ser. ICML '09. New York, NY, USA:

- Association for Computing Machinery, 2009, p. 41–48. [Online]. Available: https://doi.org/10.1145/1553374.1553380
- [22] Y. Wang, H. Chen, Q. Heng, W. Hou, Y. Fan, Z. Wu, J. Wang, M. Savvides, T. Shinozaki, B. Raj et al., "Freematch: Self-adaptive thresholding for semi-supervised learning," arXiv preprint arXiv:2205.07246, 2022.
- [23] J. Chen, C. Dun, and A. Kyrillidis, "Fast fixmatch: Faster semi-supervised learning with curriculum batch size," in 2024 IEEE International Symposium on Information Theory (ISIT), 2024, pp. 1836–1841.
- [24] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Proceedings of the european* conference on computer vision (eccv), 2018, pp. 135–152.
- [25] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12674–12684.
- [26] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. Hughes, and D. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *International conference on medical image* computing and computer-assisted intervention. Springer, 2017, pp. 408–416.
- [27] S. Xie, H. Huang, Z. Niu, L. Lin, and Y.-W. Chen, "Medfct: A frequency domain joint cnn-transformer network for semi-supervised medical image segmentation," in 2023 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2023, pp. 1913–1918.
- [28] Y. Bai, D. Chen, Q. Li, W. Shen, and Y. Wang, "Bidirectional copy-paste for semi-supervised medical image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11514–11524.
- [29] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.
- [30] H. Huang, Y. Huang, S. Xie, L. Lin, T. Ruofeng, Y.-w. Chen, Y. Li, and Y. Zheng, "Semi-supervised convolutional vision transformer with bi-level uncertainty estimation for medical image segmentation," in *Proceedings* of the 31st ACM International Conference on Multimedia, 2023, pp. 5214–5222.
- [31] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., "Segment anything," arXiv preprint arXiv:2304.02643, 2023.
- [32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.
- [33] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.
- [34] X. Liu, W. Li, and Y. Yuan, "Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation," 2024. [Online]. Available: https://arxiv.org/abs/2407.09918
- [35] J. Miao, C. Chen, K. Zhang, J. Chuai, Q. Li, and P.-A. Heng, "Cross prompting consistency with segment anything model for semi-supervised medical image segmentation," 2024. [Online]. Available: https://arxiv.org/abs/2407.05416
- [36] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv preprint arXiv:1607.06450, 2016.
- [37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference* on machine learning. pmlr, 2015, pp. 448–456.
- [38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information* processing systems, vol. 25, 2012.
- [39] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester et al., "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.
- [40] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," arXiv preprint arXiv:1902.03368, 2019.
- [41] G. Litjens, R. Toth, W. Van De Ven, C. Hoeks, S. Kerkstra, B. Van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang et al., "Evaluation of prostate segmentation algorithms for mri: the promise12 challenge," *Medical image analysis*, vol. 18, no. 2, pp. 359–373, 2014.
- [42] Z. Zhang, R. Ran, C. Tian, H. Zhou, X. Li, F. Yang, and Z. Jiao, "Self-aware and cross-sample prototypical learning for semi-supervised medical

- image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 192–201.
- [43] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.
- [44] H. Robbins and S. Monro, "A stochastic approximation method," The annals of mathematical statistics, pp. 400–407, 1951.