2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

InteractMatch: Segment Anything Model with Interact-Consistency for Semi-Supervised Medical Image Segmentation

Haohua Chang[1][0009-0007-4996-5603]

[1] Massey University, Auckland 0632, New Zealand
haohuachang@outlook.com

Abstract. The scarcity of labeled data is a significant challenge in medical image segmentation tasks. In recent years, Segment Anything Model (SAM) has gained attention as a foundational model for segmentation tasks due to its powerful zero-shot capabilities and prompt-based interactive manner. However, due to the substantial domain gap between medical and natural image data, adapting SAM to the medical domain requires a large volume of annotated medical data. Unfortunately, in medical applications, obtaining densely annotated data is both costly and challenging, particularly for rare diseases. Therefore, to efficiently fine-tune SAM, we consider utilizing Semi-Supervised Learning (SSL) to harvest knowledge from unlabeled samples. In this paper, we present InteractMatch, which consists of a Prompt Augmentation-Based Consistency (PAC) and a Cross-Model Knowledge Distillation (CKD). The PAC module effectively leverages various types of prompts from SAM to facilitate model training on unlabeled data, improving both robustness and predictive accuracy by introducing perturbations to the prompts. Additionally, CKD is introduced to align the probability distributions of the two model branches, thereby reducing discrepancies in their predictions and enhancing the output invariance of the model. Extensive experiments on two public datasets demonstrate that our InteractMatch achieves state-of-the-art performance in semi-supervised medical image segmentation task, particularly, leading a 1.93% dice score improvement on the ACDC dataset. Code is available at: https://github.com/haohua-chang/InteractMatch.

Keywords: Segment Anything Model, Semi-Supervised Learning, Medical Image Segmentation.

1 Introduction

Medical image segmentation is a critical step in accurate diagnosis and treatment planning [1], aiming to recognize target anatomical or pathological structure within images [2]. In recent times, the Segment Anything Model (SAM) [3] as an interactive foundation segmentation model has demonstrated strong generalization capabilities in various natural image segmentation tasks [4]. Nonetheless, directly applying SAM to medical image segmentation tasks has been shown to underperform compared to current state-

of-the-art medical segmentation algorithms, such as U-Net-based models and Transformer variants [5,6]. This is primarily due to the significant differences between medical data and natural images, and SAM lacks of domain-specific medical knowledge [7], which hinders its ability to associate segmented regions with meaningful semantic categories [2]. Consequently, SAM needs to be fine-tuned on medical image data to learn the relevant features.

Current approaches applying SAM to medical image segmentation mainly rely on fully-supervised training manner that require large amounts of labeled data [2,7,8,9]. However, expert annotations in medical practice are expensive and hard to collect [10], especially for some rare diseases. Compared to labeled medical image segmentation data, unlabeled data is easier to obtain and remarkably more abundant in most medical applications. To fully leverage the large volume of unlabeled medical data for efficient fine-tuning of the SAM, we aim to employ Semi-supervised Learning (SSL) to harvest the knowledge from unlabeled data for superior segmentation. In such a way, SSL allows the model to achieve performance similar to the model trained via fully-supervised learning [4], even if only using limited labeled data.

Recently, many latest works have proposed using semi-supervised methods to fine-tune the SAM [1,4,10,11]. Among these, some approaches treat the SAM as an independent component, using the pseudo-labels generated by the SAM as additional supervised signals to train the model. Other methods apply simple perturbations to the point prompts within a semi-supervised learning framework to enhance model performance. Nevertheless, these methods do not fully account for the unique characteristics of SAM's prompt mechanism. Inspired by image-based augmentation techniques, which enhance the model's discriminative capability by perturbing the input data, this paper intends to increase the perturbation of the prompts to make the model more robust to input prompts and stabilize its predictions. In such manner, the model is able to enhance its discriminative capability across different target types or regions. To this end, we propose augmenting all types of SAM prompts and devising a corresponding semi-supervised learning strategy. In addition, as our framework predicts coarse mask predictions by 2 different SAM decoders respectively, we strive to ensure that the knowledge learned by the two decoders to exhibit strong consistency, ensuring that the mask predictions across different models reflect similar information.

To achieve this goal, this study proposes an InteractMatch framework, which contains Prompt Augmentation-based Consistency (PAC) and Cross-Model Knowledge Distillation (CKD). Specifically, in PAC, the model extracts the center points and center boxes from the unprompted predictions of the unlabeled data, and these different types of prompts are treated as determined prompts. Meanwhile, data augmentation is applied to these prompts, and the augmented results are considered as ambiguous prompts. Then, the consistency loss between the ambiguous prompt predictions and the final predictions is computed, effectively utilizing the large amount of unlabeled data and SAM prompting mechanism to enhance model performance. Moreover, upon two decoder branches, the CKD regards them as a teacher and a student model respectively, and performs dynamic distillation between them to ensure the two branches producing similar predictions. Different from previous knowledge distillation method that only adopting the distillation loss between the two different segmentation models, the

proposed CKD performs the distillation between two different decoders while maintaining the encoders identical. In such way, CKD can strengthen the output consistency under different perturbations and effectively improves the model's segmentation performance. The key contributions presented in this paper are as follows:

- We propose the InteractMatch framework to harvest knowledge from unlabeled data for medical image segmentation, which comprises PAC and CKD to enhance the perturbation of prompt and improve the consistency of the predictions of the two branches, thus utilizing the large amount of unlabeled data to improve the accuracy and robustness of the model.
- In PAC, we propose determined prompting strategy and ambiguous prompting strategy to perturb different types of prompts and compute their consistency loss with the integration results, which fully utilizes the SAM model prompting mechanism to train the model effectively on the unlabeled data.
- To enforce the mask consistency among different decoders, we present CKD, which allows the model to learn a more stable probability distribution by calculating the KL scatter loss between two branches, avoiding highly biased predictions and enhancing the robustness of the model.
- Extensive experiments on two benchmarks verify that our method outperforms existing methods on the breast cancer segmentation task and the cardiac structure segmentation task, especially when labeling data is extremely limited. To be specific, using only 10 MRI slices, our method obtains a Dice improvement of more than 1.93% over various strong baselines on the cardiac structure segmentation task.

2 Related work

2.1 Segment Anything Model

In the latest years, Segment Anything Model (SAM) has demonstrated outstanding performance across segmentation tasks in various domains. Trained on 1.1 billion masks, SAM is capable of performing zero-shot segmentation across a wide range of tasks. Its prompt-based interface allows users to provide segmentation prompts, such as points, boxes, text, and masks, and interactively generates multiple possible segmentation results in real time.

Although SAM achieves performance comparable to state-of-the-art fine-tuned models in many segmentation tasks [3], it still has certain limitations, particularly in domain-specific applications. In response, various improved versions of SAM have been proposed, which can be broadly categorized into two major groups. The first category focuses on refining segmentation quality by enhancing SAM’s ability to capture finer details in target regions [12,13]. For instance, HQ-SAM [12] introduces HQ-Output Tokens to directly generate high-quality masks, which produces segmentation results with more precise boundaries and richer details while retaining strong generalization capabilities. The second category is designed to improve SAM’s generalization ability and adaptability across diverse applications [14,15]. For example, SEEM [14] incorporates a memory mechanism and multimodal fusion, enhancing the model’s

flexibility and usability. It supports a broader and more composable set of prompts, making it particularly well-suited for complex and diverse real-world scenarios.

Furthermore, since the direct application of SAM to medical image segmentation has shown suboptimal performance, several fine-tuned versions of SAM for the medical domain have been proposed. These approaches can be classified into two categories. The first category is based on Parameter-Efficient Fine-Tuning (PEFT) [2,7], which integrate domain-specific knowledge into SAM. Med-SA [7] adapts the SAM for medical applications by incorporating a spatial-depth transposition mechanism to extend 2D SAM to 3D medical images and a hyper-prompt adapter to enhance prompt-conditioned adaptation. SAMed [2] employs Low-Rank Adaptation on SAM's image encoder and mask decoder, combined with a warm-up optimization strategy, achieving state-of-the-art performance with reduced storage and deployment costs. The second category focuses on enhancing the robustness of the model [8,9]. In this context, weakly supervised self-training [9] leverages anchor-based regularization and low-rank fine-tuning to develop a task-agnostic framework, further improving SAM's robustness in medical image segmentation.

However, most of the current SAM-based methods require a large amount of labeled data to achieve satisfactory performance, which does not solve the problem of scarce labeled data in medical image analysis. To overcome this challenge, in this study, semi-supervised techniques are used to extract more information from a large amount of unlabeled data to improve the segmentation performance under limited labeling.

2.2 Semi-Supervised Learning

Supervised semantic segmentation research [16,17] has made significant progress in recent years. Despite this, these methods heavily rely on large-scale, high-quality pixel-wise annotations, which are both expensive and time-consuming to obtain, particularly for dense pixel-level labeling. To mitigate the high annotation costs, semi-supervised semantic segmentation (SSS) has been proposed, where models are trained using a small amount of labeled data alongside a large pool of unlabeled data.

Based on existing research, semi-supervised segmentation methods can be classified into two main approaches: (1) Pseudo-label-based methods [18,19]: These methods generate pseudo-labels for unlabeled images and use them to retrain the model. For instance, FixMatch [18] generates pseudo-labels from weakly augmented data and employs consistency training by applying these pseudo-labels to strongly augmented versions of the same data. ST++ [19] improves the utilization of pseudo-labels in semi-supervised learning by incorporating reliability filtering and curriculum learning, addressing issues such as error accumulation in pseudo-labels and the neglect of sample difficulty. (2) Consistency regularization-based methods [20,21,22]: These methods seek to improve model performance by enforcing consistent predictions on unlabeled data under different perturbations. Mean Teacher [20] utilizes an exponential moving average (EMA) update mechanism in a teacher-student architecture, ensuring prediction consistency between the teacher and student models as a form of regularization. CPS [21] enforces cross-consistency supervision between two networks, encouraging them to produce consistent predictions on unlabeled data and preventing overfitting to

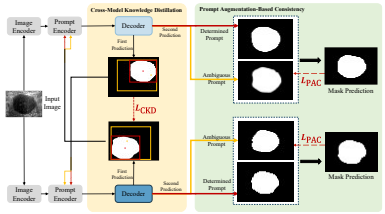


Fig. 1. The overview of InteractMatch method, which integrates PAC and CKD to enhance model training on unlabeled data through prompt-based data augmentation. The first prediction is unprompted prediction (default prompt), the second prediction is prompted prediction.

self-generated pseudo-labels.

Given the capability of semi-supervised learning to efficiently leverage information from unlabeled data, recent studies have explored the use of semi-supervised strategies to train the SAM model. Li et al. [4] generate prompts based on CNN predictions and select outputs with high consistency between CNN and SAM pseudo-labels. SemiSAM [11] integrates SAM's outputs as an additional supervision signal within the Mean Teacher framework. Nevertheless, these approaches treat SAM merely as an independent component for pseudo-label generation without considering its prompt-based characteristics, limiting model performance. It is also worth noting that CPC-SAM [1] employs a cross-prompting strategy for model training but only applies simple data augmentation to point prompts, lacking a comprehensive utilization of SAM's prompt mechanism. To address these limitations, this paper proposes to augment the data with all types of SAM prompts, design tailored semi-supervised learning strategies for efficient fine-tuning of SAM, and introduce a knowledge distillation strategy to improve the stability and accuracy of model predictions. These enhancements are expected to mitigate the challenge of limited annotated data in medical image segmentation tasks.

3 Methods

We propose the InteractMatch framework, which incorporates PAC and CKD, designed for SSL in medical image segmentation, as illustrated in Fig. 1. Considering SAM's promptable nature and its sensitivity to prompt locations in SSL, we develop PAC, which fully exploits different types of prompts. This approach incorporates both a determined prompting strategy and an ambiguous prompting strategy to enhance the model's segmentation performance under various prompt perturbations. Furthermore, we introduce CKD within a dual-branch model architecture by employing unidirectional knowledge distillation. In this manner, the model achieves greater robustness while mitigating potential negative interactions between the two branches that may arise from bidirectional knowledge distillation.

3.1 Preliminary

Segment Anything Model (SAM). SAM [3] serves as a foundation model for segmentation tasks, designed for powerful zero-shot segmentation capability. It employs a Vision Transformer (ViT) as the image encoder and integrates a Prompt Encoder and a Mask Decoder to enable interactive segmentation. Given a labeled dataset $S = \{(I_i, M_i)\}_{i=1}^{N}$, where $I_i \in \mathrm{R}^{(\mathrm{H}\times\mathrm{W}\times 3)}$ represents the i-th RGB image with a resolution of $H \times W$, and $M_i \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}}$ denotes the corresponding ground-truth segmentation mask, the image encoder of SAM is denoted as E_I, the prompt encoder as E_P, and the mask decoder as D. For a given image $I \in S$, the image embedding I_{emb} is obtained using the image encoder: $I_{\text{emb}} = E_I(I; \theta_I)$, where θ_I represents the parameters of the image encoder. Simultaneously, a prompt is processed by the prompt encoder to generate the prompt embedding P_{emb}: $P_{\text{emb}} = E_P(P; \theta_P)$, where prompts can be of three types: points, bounding boxes, or masks. Ultimately, the mask decoder produces the predicted segmentation mask p: $p = D(I_{\text{emb}}, P_{\text{emb}}; \theta_D)$. The entire data flow of SAM can be further simplified as: $p = D\big(E_I(I), E_P(P)\big)$.

Overview of InteractMatch. The InteractMatch framework integrates SSL strategies with the promptable capabilities of SAM, enabling joint training using a small set of labeled data $S = \{(I_i, M_i)\}_{i=1}^{N}$ and a large set of unlabeled data $U = \{(I_i)\}_{i=N+1}^{N+M}$. To ensure high-quality predictions in the early training stage and leverage SAM's strong few-shot learning capability, we first fine-tune SAM on all labeled data. Building on this, we train the model on unlabeled data using the PAC strategy. Owing to the demonstrated efficacy of cross-pseudo supervision [21] in improving SSL outcomes, we adopt a dual-branch structure, following the design in [1]. To be more specific, our model employs two mask decoders with the same structure but different initializations, D_1 and D_2, which share a common image encoder E_I and prompt encoder E_P. Each module retains the same functionality as in the original SAM [3]. Taking Branch 1 as an example, for a given input image I, the model first generates an unprompted prediction p_1: $p_1 = D_1\big(E_I(I), E_P(P_{\text{default}})\big)$. Subsequently, we employ the prompts P_{d2} and P_{a2}, generated by the PAC method in branch 2, to perform prompted predictions, yielding $\hat{p}_{1,d}$ and $\hat{p}_{1,a}$: $\hat{p}_{1,d} = D_1\big(E_I(I), E_P(P_{\text{d2}})\big)$, $\hat{p}_{1,a} = D_1\big(E_I(I), E_P(P_{\text{a2}})\big)$. The final output of branch 1, $\hat{p}_1$, is the mean of $\hat{p}_{1,d}$ and $\hat{p}_{1,a}$. Branch 2 follows the same process, generating p_2, $\hat{p}_{2,d}$, $\hat{p}_{2,a}$, and $\hat{p}_2$. In addition, to further enhance segmentation stability, we apply CKD to the unprompted predictions p_1 and p_2 from both branches.

3.2 Prompt Augmentation-Based Consistency

PAC consists of the determined prompting strategy and the ambiguous prompting strategy and performs two predictions. The first prediction is unprompted prediction, where the determined prompting strategy and ambiguous prompting strategy are applied to the generated coarse mask to obtain the determined prompts and ambiguous prompts. Subsequently, the prompts generated by both strategies are input back into the model to generate the prompted predictions. Finally, the prompted predictions obtained from both strategies are averaged to form pseudo-labels, which are used to calculate the

consistency loss with respect to the predictions generated under the ambiguous prompting strategy.

Determined prompting strategy. To generate more accurate prompt information, we design the determined prompting to extract the center point of the mask's shape and the smallest enclosing box from the coarse mask generated in the first prediction. Given that the coarse mask may contain small noisy regions, we select the largest connected component within the coarse mask as the Region of Interest (ROI), and from this ROI, we generate the aforementioned center point and bounding box. By generating these two distinct and more accurate types of prompts as location cues for the SAM model, this strategy helps guide the model toward producing more precise prediction results.

Ambiguous prompting strategy. To generate vague or relatively inaccurate prompt information, the proposed ambiguous prompting selects a random point within the ROI of the coarse mask from the first prediction as a point prompt, and generates a box larger than the ROI as an enlarged offset box. Importantly, this box is created by randomly shifting and enlarging the box generated in the determined prompting strategy along various directions. To further enhance this strategy, the box is designed to always enclose the entire ROI, as this strategy aims to avoid producing excessively inaccurate location prompts that could negatively affect the model's predictions. By generating these two types of vague prompts, the model can better learn the consistency of outputs under different prompting conditions.

Loss function. Due to the use of a dual-branch architecture, where the loss calculation process for both branches is identical, we explain how the PAC loss is computed using Branch 1 (D_1) as an example. First, the prompt used by Branch 1 is generated by Branch 2. Branch 2 uses the default prompt P_{default} to generate the unprompted prediction p_2, where following [2], P_{default} is fine-tuned during training for automatic segmentation and also employed during inference. Then, the determined prompting strategy and ambiguous prompting strategy are applied to p_2, generating the determined prompt P_{d2} and the ambiguous prompt P_{a2}, respectively. These prompts are input into D_1 to generate the prompted predictions $\hat{p}_{1,d}$ and $\hat{p}_{1,a}$. We then integrate the prompted prediction results from the two strategies, $\hat{p}_1 = (\hat{p}_{1,d} + \hat{p}_{1,a})/2$, which provides a more robust result as pseudo-labels to guide the model in learning the correct features and effectively improving its performance. Thereafter, the PAC loss between $\hat{p}_{1,a}$ and $\hat{p}_1$ is calculated. Similarly, we generate the prompted predictions $\hat{p}_{2,d}$, $\hat{p}_{2,a}$, and their integrated result $\hat{p}_2$ through D_2. The PAC loss L_{PAC} is then applied to both branches, as shown in equation (1):

$$L_{\text{PAC}} = \frac{1}{2}\left[L_{\text{dice}}(\hat{p}_{1,a}, \hat{p}_1) + L_{\text{ce}}(\hat{p}_{1,a}, \hat{p}_1)\right] + \frac{1}{2}\left[L_{\text{dice}}(\hat{p}_{2,a}, \hat{p}_2) + L_{\text{ce}}(\hat{p}_{2,a}, \hat{p}_2)\right]. \quad (1)$$

By integrating the prediction results from both prompting strategies, the model will generate more accurate predictions. Calculating the PAC loss between these predictions and the predictions obtained under ambiguous prompting will enhance the model's robustness to different prompt perturbations, improving its ability to perform accurate segmentation.

3.3 Cross-Model Knowledge Distillation

To further enhance the model's robustness and accuracy, we introduce CKD across the two branches. By calculating the KL divergence loss between the prediction results of the two branches in the unprompted prediction, branch 1 is treated as the teacher model for unidirectional knowledge distillation, allowing branch 2 to learn a similar probability distribution as branch 1. This method treats the segmentation task as a series of independent pixel classification problems [23], and uses knowledge distillation to align the class probabilities for each pixel generated by branch 2 with the class probabilities (soft labels) generated by branch 1. The CKD loss L_{CKD} is defined in equation (2):

$$L_{\mathrm{CKD}} = L_{\mathrm{KL}}(p_2, p_1), \tag{2}$$

where L_{KL} represents the KL divergence loss for segmentation tasks, as calculated following the method described in [23], and is given by equation (3):

$$L_{\mathrm{KL}} = \frac{1}{W \times H} \sum_{i \in R} KL(q_i^2 \parallel q_i^1), \tag{3}$$

where W, H represent the width and height of the image, respectively, and R denotes the total number of pixels in the image. Normalization is applied to balance the loss calculation, ensuring that the loss function does not influence the gradient update due to differences in pixel count. q_i^2 denotes the class probability distribution of the i-th pixel generated by branch 2, and q_i^1 denotes the class probability distribution of the i-th pixel generated by branch 1.

The CKD introduced in this study imposes a consistency constraint between the two branches, helping the model avoid predictions with significant deviations from the ground truth during training. To be more precise, if there is a large discrepancy between the predictions of the two branches, the CKD method gradually aligns the predictions of branch 2 with those of branch 1, consequently reducing the occurrence of highly divergent predictions and improving the model's robustness.

3.4 Optimization and Inference

Optimization. The total loss for training the InteractMatch architecture is a combination of the supervised losses L_{sup1} and L_{sup2} on labeled data, the cross-supervision loss L_{cross} and PAC loss L_{PAC} on unlabeled data, and the CKD loss L_{CKD} on all data, as shown in equation (4).

$$L_{\mathrm{total}} = L_{\mathrm{sup1}} + L_{\mathrm{sup2}} + \lambda_1 L_{\mathrm{CKD}} + \lambda_2 L_{\mathrm{cross}} + \lambda_3 L_{\mathrm{PAC}}. \tag{4}$$

Firstly, proposed method fine-tunes the model using labeled data before training with a large amount of unlabeled data. Thus, we compute the loss between the unprompted prediction and the corresponding labels, as shown in equation (5).

$$L_{\mathrm{sup1}} = L_s(p_1, \mathbf{y}) + L_s(p_2, \mathbf{y}), \tag{5}$$

where L_s represents the combination of dice score loss and Cross-Entropy (CE) loss. This loss function enables the model to quickly learn effective features from the data, ensuring the prediction quality during the early stages of the unsupervised learning phase.

Secondly, to ensure that the prompted predictions progressively approach the ground truth, we also compute the supervised loss L_{sup2} for the predictions of both branches under the determined prompting strategy and ambiguous prompting strategy, as shown in equation (6).

$$L_{\text{sup2}} = L_s(p_{1,d}, \mathbf{y}) + L_s(p_{1,a}, \mathbf{y}) + L_s(p_{2,d}, \mathbf{y}) + L_s(p_{2,a}, \mathbf{y}). \quad (6)$$

Finally, we calculate the loss between the unprompted prediction p_1 from branch 1 and the prompted prediction $\hat{p}_2$ from branch 2, as well as the loss between p_2 and $\hat{p}_1$. Cross-branch supervision alleviates the confirmation bias issue inherent in self-supervision within the same branch [1]. The calculation of the cross-supervision loss L_{cross} is shown in equation (7).

$$L_{\text{cross}} = \frac{1}{2}[L_{\text{dice}}(p_1, \hat{p}_2) + L_{\text{ce}}(p_1, \hat{p}_2)] + \frac{1}{2}[L_{\text{dice}}(p_2, \hat{p}_1) + L_{\text{ce}}(p_2, \hat{p}_1)]. \quad (7)$$

Inference. During inference, the model uses the default prompt for prediction, generates segmentation masks using both branches, and take the average of these masks as the final prediction result.

4 Experiments

4.1 Datasets

This paper evaluates the InteractMatch method on two public medical segmentation datasets: the ACDC dataset [24] and the BUSI dataset [25].

ACDC. The Automatic Cardiac Diagnosis Challenge (ACDC) is a dataset for cardiac MRI (CMR) evaluation with the goal of segmenting the myocardium, the right ventricle cavity and the left ventricle cavity. The dataset contains a total of 200 cine MRI scans of 100 patients at end-diastole (ED) and end-systole (ES) of the heart. Following [31,32], the dataset will be randomly split at the patient level, where 70 patients will be used for training, 10 for validation, and 20 for testing.

BUSI. The dataset is medical images of breast cancer and includes a total of 780 ultrasound images. These images are categorized into three categories i.e. normal, benign and malignant. In this article, we will randomly split the data of benign and malignant categories and finally obtain 431, 86 and 130 images for training, validation and testing respectively.

4.2 Implementation Details and Evaluation Metrics

Implementation Details. The InteractMatch method is implemented using PyTorch and trained on a single NVIDIA 3090 GPU. The training process is conducted on the ACDC and BUSI datasets under approximately identical settings. In detail, the ViT-base version of the SAM is utilized, with LoRA [26] is applied to the query and value heads within each transformer block of E, using a rank $r = 4$, while all parameters in

E_P and D_1, D_2 are optimized through standard backpropagation, following the approach in [2]. Following [1], the pre-trained weights are loaded for both the image encoder and prompt encoder, whereas the two decoders are randomly initialized. Additionally, the output resolution of the mask decoders is increased using a progressive upsampling strategy, as proposed in [27]. Input images are resized to 512×512 and normalized within the range [0,1]. Data augmentation following [1]. The adapted SAM is trained for 8000 epochs using the AdamW optimizer. A warmup period of 5,000 iterations is employed, followed by an exponential learning rate decay, consistent with [2]. The maximum learning rate is set to 0.001 and the hyperparameters λ_2 is empirically set to 0.4. The batch size is set to 6 for the BUSI dataset and 12 for the ACDC dataset, with each batch comprising an equal proportion of labeled and unlabeled data.
Evaluation Metrics. The performance of the proposed segmentation method is assessed using four evaluation metrics: the Dice Similarity Coefficient (DSC), Jaccard (JC), 95th percentile Hausdorff Distance (95HD), and Average Surface Distance (ASD). The DSC quantifies the similarity between the predicted segmentation and the ground truth by measuring the overlap between the two regions. A higher DSC value indicates a greater correspondence between the predicted and actual segmentations. The JC, also referred to as the Intersection over Union (IoU), provides an alternative measure of overlap; however, it is a more stringent metric as it is directly based on the IoU calculation. The 95HD evaluates the 95th percentile of the distances from the predicted segmentation boundary to the ground truth boundary, with lower values indicating a closer alignment between the two. The ASD computes the mean distance between the predicted and actual segmentation boundaries, serving as a measure of overall segmentation error. Both DSC and JC are expressed as percentages, while 95HD and ASD are measured in pixels for the BUSI dataset and in millimeters for the ACDC dataset, as the BUSI dataset does not provide resolution information.

4.3 Comparison with the State-of-the-art Networks

We evaluate the proposed method against state-of-the-art (SOTA) SSL methods, including UAMT [28], CPS [21], URPC [29], MC-Net+ [30], DCNet [31], BCP [32], UniMatch [22], SemiSAM [11], and CPC-SAM[1]. To further evaluate performance, we compare our method with fully supervised baselines trained on labeled data (U-Net [33]) and the zero-shot SAM using a center-point prompt derived from ground truth labels as in [34], denoted as "SAM-point". To ensure a fair comparison, we selected the optimal results of these methods on the ACDC and BUSI test sets. As shown in Table 1, our method demonstrates a significant performance improvement over the other SSL approaches on the ACDC dataset. In particular, across different amounts of labeled data, our method leads to improvements of 1.93% and 0.87% DSC scores over the second-best competitor, while JC scores surpass others by 2.6% and 1.29%, respectively. Besides, our approach achieves substantial performance gains in HD95 and ASD, reducing their values by 3.19 mm and 1.35 mm, respectively, in the single labeled sample experiment compared to the second-best approach. Meanwhile, InteractMatch attains competitive sensitivity and specificity scores of 88.31% and 99.86%, respectively, in a single labeled data experiment. Moreover, to assess the robustness of our method, we

conducted experiments with random seeds of 42, 542, and 1337. The average DSC and its standard deviation obtained from three different seeds were 86.6% and 0.0000403, respectively. The experimental comparison results in Table 2 demonstrate that our method consistently outperforms competing approaches across all evaluation metrics and labeled data ratios on the BUSI dataset. Notably, under the SSL setting, our method achieves DSC improvements of at least 1.02% and 1.76% in experiments with 10 and 20 labeled samples, respectively.

Table 1. Comparison with state-of-the-art (SOTA) methods on the ACDC dataset. The column "#Lab" indicates the number of labeled samples and the total number of training samples, respectively.

Method	ACDC				
	#lab	DSC↑	JC↑	HD95↓	ASD↓
U-Net [33]	70/70	91.53	84.77	4.23	1.11
SAM-point (MIA'23) [34]	0/70	62.88	49.53	20.46	7.07
U-Net [33]	1/70	29.37	20.53	107.51	52.84
SAMed [2]	1/70	75.01	61.53	28.99	9.13
UAMT(MICCAI'19) [28]	1/70	29.14	20.14	107.69	53.58
CPS(CVPR'21) [21]	1/70	30.46	21.00	95.74	45.48
URPC(MIA'22) [29]	1/70	31.00	20.81	123.03	59.94
MC-Net+(MIA'22) [30]	1/70	38.84	28.58	62.21	30.67
DCNet(MICCAI'23) [31]	1/70	41.13	31.6 1	56.16	24.71
BCP(CVPR'23) [32]	1/70	68.39	56.8	50.9	21.99
UniMatch(CVPR'23) [22]	1/70	84.47	74.25	15.36	4.57
SemiSAM [11]	1/70	34.18	23.96	100.75	47.03
CPC-SAM [1]	1/70	85.56	75.74	9.19	2.84
InteractMatch (ours)	1/70	**87.49**	**78.34**	**6.00**	**1.49**
U-Net [33]	3/70	45.95	35.96	71.11	32.47
SAMed [2]	3/70	83.04	71.98	14.93	4.05
UAMT(MICCAI'19) [28]	3/70	56.67	45.93	15.06	45.24
CPS(CVPR'21) [21]	3/70	56.87	46.88	20.18	2.91
URPC(MIA'22) [29]	3/70	55.98	44.75	40.47	14.13
MC-Net+(MIA'22) [30]	3/70	65.37	54.18	27.64	6.32
DCNet(MICCAI'23) [31]	3/70	72.21	62.27	26.50	10.59
BCP(CVPR'23) [32]	3/70	87.57	78.58	8.68	2.30
UniMatch(CVPR'23) [22]	3/70	87.31	78.20	8.62	2.74
SemiSAM [11]	3/70	51.01	39.45	70.13	28.26
CPC-SAM [1]	3/70	87.95	79.01	5.80	1.54
InteractMatch (ours)	3/70	**88.82**	**80.30**	**5.53**	**1.34**

Regarding computational efficiency, our approach requires 22,630 MB of GPU memory, slightly higher than the 22,497 MB occupied by CPC-SAM. The average inference time per slice on the ACDC dataset is 0.0336 s for our method and 0.0318 s for CPC-SAM. These findings demonstrate that the proposed framework incurs negligible

additional computational overhead while yielding notable enhancements in segmentation accuracy.

Table 2. Comparison with state-of-the-art (SOTA) methods on the BUSI dataset. The column "#Lab" indicates the number of labeled samples and the total number of training samples, respectively.

Method	BUSI				
	#lab	DSC↑	JC↑	HD95↓	ASD↓
U-Net [33]	431/431	77.19	68.29	75.03	31.21
SAM-point(MIA'23) [34]	0/431	52.99	44.51	168.26	91.78
U-Net [33]	10/431	31.63	24.52	159.49	63.43
SAMed [2]	10/431	65.09	54.78	119.75	47.84
UAMT(MICCAI'19) [28]	10/431	40.93	30.96	175.31	76.51
CPS(CVPR'21) [21]	10/431	32.92	25.70	144.92	50.54
URPC(MIA'22) [29]	10/431	32.16	24.75	151.59	64.97
MC-Net+(MIA'22) [30]	10/431	36.24	27.45	167.91	71.80
DCNet(MICCAI'23) [31]	10/431	42.14	32.11	154.39	64.21
BCP(CVPR'23) [32]	10/431	61.81	51.12	112.91	38.15
UniMatch(CVPR'23) [22]	10/431	60.98	49.85	109.79	47.50
SemiSAM [11]	10/431	43.43	32.48	177.30	84.46
CPC-SAM [1]	10/431	71.20	61.15	100.22	37.86
InteractMatch (ours)	10/431	**72.22**	**63.10**	**85.78**	**33.04**
U-Net [33]	20/431	44.22	34.73	160.04	69.52
SAMed [2]	20/431	67.28	57.55	107.31	49.70
UAMT(MICCAI'19) [28]	20/431	45.83	35.84	163.53	80.92
CPS(CVPR'21) [21]	20/431	46.74	37.61	142.73	56.70
URPC(MIA'22) [29]	20/431	45.26	35.51	173.11	73.47
MC-Net+(MIA'22) [30]	20/431	47.29	33.00	183.14	84.53
DCNet(MICCAI'23) [31]	20/431	56.87	46.60	130.31	56.14
BCP(CVPR'23) [32]	20/431	65.54	56.05	93.07	39.09
UniMatch(CVPR'23) [22]	20/431	62.47	51.48	100.73	45.88
SemiSAM [11]	20/431	50.09	38.63	170.42	77.85
CPC-SAM [1]	20/431	72.41	62.72	96.26	40.93
InteractMatch (ours)	20/431	**74.17**	**65.21**	**74.30**	**28.78**

The comprehensive performance improvements of the InteractMatch approach on the ACDC and BUSI datasets demonstrate its enhanced adaptability to diverse anatomical structures in medical imaging. Overall, the robustness and superior performance of our method demonstrate the effectiveness of the InteractMatch approach.

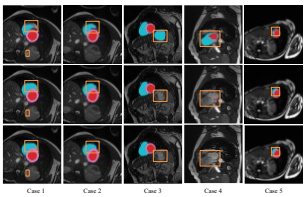


Fig. 2. Visualizations of CPC-SAM method (top), InteractMatch method (middle) and corresponding Ground Truth (bottom) in 5 cases on the ACDC dataset with 1 labeled patient. The red, blue, and pink regions respectively correspond to the left ventricular cavity, right ventricular cavity, and myocardium. The orange-colored boxes highlight the key differences between the various methods and the ground truth.

To further validate the advantages of our method, qualitative comparisons are presented in Fig. 2. As depicted in the figure, our approach exhibits a marked improvement in robustness when dealing with noise in small regions, as evidenced by the boxed area

Table 3. Ablation studies of different components of our method on the ACDC dataset.

CKD	PAC	DSC↑	JC↑	HD95↓	ASD↓
✓	✓	**87.49**	**78.34**	**6.00**	**1.49**
✓		86.97	77.94	9.14	2.70
	✓	85.11	75.07	10.76	3.47

located at the bottom of Case 1. In particular, it achieves more accurate delineation of target regions, as evidenced by the boxed areas at the top of Case 1 and in Case 2. Besides, the InteractMatch approach effectively reduces both false-positive predictions (Cases 3 and 4) and false-negative predictions (Case 5). It is particularly worth emphasizing that minimizing false negatives is critical, as failing to detect abnormal tissues in medical image analysis can have severe consequences.

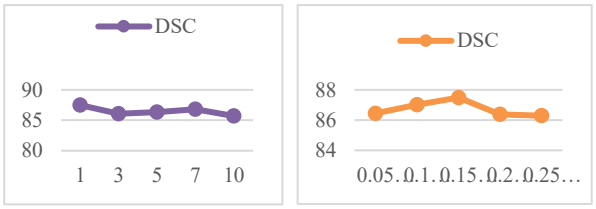
4.4 Ablation Experiments

Table 3 shows the ablation study of the two key components of the proposed method on the ACDC dataset. It is clear that 4 evaluation metrics of the model deteriorate when either of the two components is removed. To be specific, with the removal of the PAC (Row 2), i.e., the prompted prediction of both branches using only the predictions generated by the center point prompt for cross-supervision and using the CDK, the DSC of the model decreases by 0.52%. In addition, after the proposed method removes the CKD component (Row 3), i.e., it no longer imposes constraints on the unprompted prediction between two branches, the DSC of the model exhibits a reduction of 2.38%. This indicates that both modules in InteractMatch play a crucial role in enhancing the model's segmentation performance.

4.5 Parameter Analysis

This section analyzes key hyperparameters in the InteractMatch framework. More specifically, multiple experiments were conducted to investigate the impact of the number of random points and different bounding boxes in the ambiguous prompting strategy, the weight of the PAC loss, and the weight of the CKD loss. The effects of these parameters on the final performance are systematically examined.

Random point number. The number of random points in the ambiguous prompting strategy provides varying degrees of position information to the model. As shown in Fig. 3(a), the DSC exhibits a general downward trend as the number of random point prompts increases. This decline may be attributed to the fact that an increased number of random point prompts provides more precise position cues, thereby reducing the ambiguity intended by the strategy and diminishing the effectiveness of the PAC method. Another potential reason is that some random points may fall outside the ground truth regions in the unlabeled data, introducing incorrect position information that adversely impacts model performance.



(a) (b)

Fig. 3. DSC for different number of random points (a) and different boxes (b).

Table 4. The model's performance under varying PAC loss weights λ_3.

λ_3	DSC↑	JC↑	HD95↓	ASD↓
0.025	86.31	76.53	8.02	2.08
0.05	**87.49**	**78.34**	**6.00**	**1.49**
0.1	86.63	77.04	7.39	2.17
0.2	85.02	74.65	9.94	2.83
0.4	84.00	73.44	19.75	5.61

Table 5. The impact of CKD loss weight λ_1 on model performance.

λ_1	DSC↑	JC↑	HD95↓	ASD↓
5	87.01	77.60	**5.56**	**1.46**
2	86.32	76.57	7.15	1.79
1	**87.49**	**78.34**	6.00	1.49
0.5	86.02	76.33	11.11	3.26
0.1	83.61	72.98	22.89	6.71

Size and offset of the box. Under the ambiguous prompting strategy, larger bounding boxes cover a broader area, resulting in more ambiguous location cues provided to the model. Conversely, smaller bounding boxes offer more precise localization information. To ensure that the expanded and shifted bounding box consistently encloses the entire ROI, the box size is directly proportional to the offset. The size and offset of the box is directly represented by the offset value; for example, an offset of 0.05 corresponds to a scaling factor of 1.1. Fig. 3(b) illustrates the impact of different box sizes and offset values on model performance. Experimental results indicate that both excessively large and overly small bounding boxes and offsets lead to performance degradation. More precisely, large offsets (e.g., 0.2 and 0.25) produce excessively ambiguous

position prompt, impairing model performance. In contrast, small offsets (e.g., 0.05) yield bounding boxes that provide nearly identical localization information to those generated by the determined prompting strategy, failing to leverage the benefits of the PAC method.

The weight of PAC loss. Table 4 presents the impact of different PAC loss weights on model performance. Since the PAC loss is computed on unlabeled data, an excessively large weight may cause the model to focus on learning irrelevant features, leading to deviations from the ground truth (e.g., when $\lambda_3 = 0.2$ and $\lambda_3 = 0.4$). Conversely, an overly small weight ($\lambda_3 = 0.025$) hinders the effectiveness of the proposed PAC method in improving model performance. Therefore, we adopt $\lambda_3 = 0.05$ as the final configuration, as it yields the best performance.

The weight of CKD loss. The CKD loss weight reflects the strength of the constraint imposed on the two branches, with an excessively small λ_1 (e.g., $\lambda_1 = 0.1$) failing to activate the CKD mechanism, as shown in Table **5**. It is worth mentioning that even with a relatively large loss weight ($\lambda_1 = 5$), the CKD method remains effective in enhancing model performance. The final value of $\lambda_1 = 1$ is selected for our approach, as it achieves the most favorable overall performance.

5 Conclusion

In this work, we present InteractMatch framework to effectively leverage SSL scheme for fine-tuning SAM on medical image segmentation task. This architecture leverages the PAC module to perturb the spatial information fed into SAM and the CKD module to enhance prediction consistency between the two branches, efficiently extracting knowledge embedded within extensive unlabeled data to improve the model's robustness and accuracy. Extensive experiments on two publicly available datasets demonstrate that our method can effectively utilize unlabeled data to improve model performance, even when labeled data is severely limited. In future work, we will explore additional perturbation strategies to further enhance the robustness of the model's outputs.

References

1. Miao, J., Chen, C., Zhang, K., Chuai, J., Li, Q., Heng, P.A.: Cross prompting consistency with segment anything model for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 167–177. Springer, Cham (2024).
2. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023).
3. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Girshick, R.: Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026. IEEE, Paris (2023).

4. Li, N., Xiong, L., Qiu, W., Pan, Y., Luo, Y., Zhang, Y.: Segment anything model for semi-supervised medical image segmentation via selecting reliable pseudo-labels. In: International Conference on Neural Information Processing, pp. 138–149. Springer, Singapore (2023).
5. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-DenseUNet: Hybrid densely connected U-Net for liver and tumor segmentation from CT volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018).
6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Zhou, Y.: TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021).
7. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Jin, Y.: Medical SAM adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023).
8. Cheng, Z., Wei, Q., Zhu, H., Wang, Y., Qu, L., Shao, W., Zhou, Y.: Unleashing the potential of SAM for medical adaptation via hierarchical decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3511–3522. IEEE (2024).
9. Zhang, H., Su, Y., Xu, X., Jia, K.: Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23385–23395 (2024).
10. Zhang, Y., Zhou, T., Wang, S., Wu, Y., Gu, P., Chen, D.Z.: SAMDSK: Combining segment anything model with domain-specific knowledge for semi-supervised learning in medical image segmentation. arXiv preprint arXiv:2308.13759 (2023).
11. Zhang, Y., Yang, J., Liu, Y., Cheng, Y., Qi, Y.: SemiSAM: Enhancing semi-supervised medical image segmentation via SAM-assisted consistency regularization. In: International Conference on Bioinformatics and Biomedicine, pp. 3982–3986. IEEE (2024).
12. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F.: Segment anything in high quality. Advances in Neural Information Processing Systems **36**, 29914–29934 (2023).
13. Wei, Z., Chen, P., Yu, X., Li, G., Jiao, J., Han, Z.: Semantic-aware SAM for point-prompted instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3585–3594 (2024).
14. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36**, 19769–19782 (2023).
15. Sun, Y., Chen, J., Zhang, S., Zhang, X., Chen, Q., Zhang, G., Li, Z.: VRP-SAM: SAM with visual reference prompt. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23565–23574 (2024).
16. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017).
17. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 801–818. Springer, Cham (2018).
18. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Li, C.L.: FixMatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems **33**, 596–608 (2020).

19. Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y.: St++: Make self-training work better for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4268–4277 (2022).
20. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in Neural Information Processing Systems **30** (2017).
21. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2613–2622 (2021).
22. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7236–7246 (2023).
23. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2604–2613 (2019).
24. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P. A., Jodoin, P. M.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018).
25. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020).
26. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021).
27. Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Li, Q.: Ma-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation. Medical Image Analysis **98**, 103310 (2024).
28. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 605–613. Springer, Shenzhen (2019).
29. Luo, X., Wang, G., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S.: Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. Medical Image Analysis **80**, 102517 (2022).
30. Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J.: Mutual consistency learning for semi-supervised medical image segmentation. Medical Image Analysis **81**, 102530 (2022).
31. Chen, F., Fei, J., Chen, Y., Huang, C.: Decoupled consistency for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 551–561. Springer, Cham (2023).
32. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11514–11524. IEEE (2023).
33. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer, Munich (2015).
34. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Medical Image Analysis **89**, 102918 (2023).