# GapMatch: Bridging Instance and Model Perturbations for Enhanced Semi-Supervised Medical Image Segmentation

**Wei Huang**[1], **Lei Zhang**[1*], **Zizhou Wang**[2], **Yan Wang**[2]

[1]College of Computer Science, Sichuan University, China
[2]Institute of High Performance Computing, A*STAR, Singapore
weihuang@stu.scu.edu.cn, leizhang@scu.edu.cn, {wang_zizhou, wangyan}@ihpc.a-star.edu.sg

## Abstract

Medical image segmentation provides detailed understanding and aids in diagnosis, treatment planning, and monitoring of diseases. Due to the high cost of acquiring labeled data in the field of medical image analysis, semi-supervised segmentation methods have garnered increasing attention. Benefiting from their simplicity and effectiveness, consistency regularization-based methods have emerged as a significant research focus by utilizing perturbations. However, existing methods typically consider perturbation strategies from only a single perspective: either instance perturbation or model perturbation, thus ignoring the potential benefit of effectively combining both. In response, we propose a unified perturbation framework named GapMatch, which bridges instance and model perturbations to broaden the perturbation space and employs dual perturbation to impose consistency regularization on the model. Specifically, GapMatch involves using instance perturbation to update the decision boundary and model perturbation to further optimize it. These two steps mutually reinforce each other in an iterative manner, effectively pushing the decision boundary towards low-density regions while maximizing the class margin. Extensive experimental results on two popular medical image benchmarks demonstrate the effectiveness and generality of the proposed method.

## Introduction

In the field of medical imaging, the accurate segmentation of anatomical structures and pathological regions plays a critical role in diagnostic processes, treatment planning, and outcome prediction (Wang et al. 2021b; Tajbakhsh et al. 2020; Wang et al. 2022c). Deep learning models, particularly Convolutional Neural Networks (CNNs), have achieved state-of-the-art performance in medical image segmentation tasks (Isensee et al. 2021). However, their success is largely contingent upon the availability of large annotated datasets, which are especially challenging to acquire in the medical domain due to the need for expert knowledge and the privacy concerns associated with patient data. Therefore, increasing attention is being focused on semi-supervised learning (SSL) methods, which allow for the effective use of both limited labeled data and the typically more abundant unlabeled data (Wu et al. 2022b; You et al. 2023b; Bai et al. 2023; You et al. 2023a).

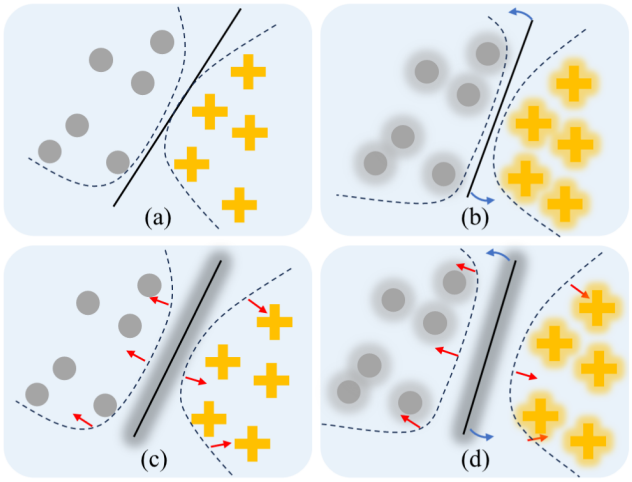


Figure 1: Illustration of the change of decision boundary of the previous methods and our proposed method. (a) Initial decision boundary and sample distribution; (b) Decision boundary update with instance perturbation-based consistency; (c) Sample distribution compression with model perturbation-based consistency; (d) (Ours) Combined instance and model perturbation based consistency regularization for decision boundary update and sample distribution compression.

Currently, the majority semi-supervised segmentation methods can be divided into two main categories: entropy minimization (Yu et al. 2019) and consistency regularization (Wang et al. 2021a; Zhang et al. 2017; Miyato et al. 2018; Liu and Zheng 2022; Grandvalet and Bengio 2004; Huang et al. 2024). Entropy minimization methods are based on the clustering assumption, which posits that the clusters should be compact within each class. It encourages the model to make confident predictions, which indirectly guides the decision boundary towards the low-density regions. Consistency regularization methods are based on the smoothness assumption (Van Engelen and Hoos 2020), which posits that small perturbations should not produce the

*Corresponding author: Lei Zhang

obvious deviations of the corresponding outputs (Laine and Aila 2016). Consistency regularization encourages the decision boundary to avoid high-density regions and instead lie in low-density regions by enforcing the model to maintain consistent predictions under small perturbations. Benefiting from its simplicity and effectiveness, consistency regularization-based methods have emerged as the mainstream in semi-supervised segmentation tasks currently.

The key aspect of the consistency regularization method lies in how perturbations are designed and implemented. Instance perturbations are a common approach employed for realizing consistency regularization. Based on the assumption that small perturbations to instances do not significantly affect predictions, perturbations such as random noise, data augmentation, and dropout are applied to the image (in input space) or the features (in feature space). These methods enforce perturbed versions of instances to maintain consistent predictions with the original ones, thus forming an avoidance zone around the instances to drive decision boundary, as shown in Fig. 1(b). Additionally, some methods focus on another orthogonal direction: model perturbations, positing that predictions should maintain consistency across varying parameter settings. In practice, they often utilize multi-branch consistency, multi-scale consistency, multi-task consistency, and consistency between teacher model and student model to facilitate model learning. As illustrated in Fig. 1(c), they can essentially be viewed as perturbations to the decision boundary, effectively compressing the distribution of instances and thereby facilitating model learning.

Although these methods can enhance the performance of semi-supervised medical image segmentation to a certain extent, they still have two shortcomings: First, while instance perturbation-based methods have been thoroughly explored, current model perturbation-based methods are inherently empirical and their performance is profoundly dependent on stochasticity. Second, most of these methods only consider perturbation strategies from a single perspective, either instance or model perspective, ignoring the combination of them and their synergistic effects. Yet, when effectively combined, these strategies can potentially improve performance due to their complementary. To this end, we propose a unified framework named GapMatch, that leverages the complementary strengths of consistency regularization from instance and model perturbation, thereby boosting overall performance.

Specifically, GapMatch first employs instance perturbation-based consistency regularization to determine the tentative direction for updating the model. Subsequently, we introduce a model perturbation named Gradient-based Adversarial Perturbation (GAP), which perturbs the model according to the determined direction from instance perturbation. Finally, GapMatch further encourages the model to maintain consistent predictions under GAP. Instance perturbation helps determine the initial position of the decision boundary while model perturbation optimizes its position. By leveraging their complementary and synergistic effects, GapMatch achieves dual optimization of the decision boundary. This ensures that the final decision boundary maintains a sufficient margin, thereby enhancing the model's overall performance and robustness. Extensive experiments conducted on two widely recognized semi-supervised medical segmentation benchmarks have conclusively demonstrated that GapMatch consistently outperforms current state-of-the-art methods, particularly when the label ratio is extremely low (1%). Furthermore, GAP is designed to be plug-and-play, enabling various SSL methods to integrate with GAP and benefit from its performance enhancements. The experimental results also confirm the generalizability of GAP.

In summary, the main contributions of this study are summarized as follows:

- We propose a framework named GapMatch, which effectively leverages both instance and model perturbations for semi-supervised medical image segmentation, demonstrating that collaboratively leveraging them is beneficial for enhancing performance.
- We propose a novel model perturbation named GAP, to extend the perturbation scope in semi-supervised medical image segmentation. GAP is a plug-and-play method that can be easily integrated with existing SSL methods.
- Extensive evaluations and consistent performance gains demonstrate the effectiveness of GapMatch. Meanwhile, multiple existing SSL methods achieve performance enhancements by incorporating GAP, demonstrating its strong generalizability.

## Related Work

### Semi-supervised Learning

Semi-supervised learning (SSL) is an effective approach that utilizes a large amount of unlabeled data along with a limited amount of labeled data. Currently, the dominant semi-supervised segmentation methods can be divided into two main categories: entropy minimization (Yu et al. 2019) and consistency regularization (Wang et al. 2021a; Zhang et al. 2017; Miyato et al. 2018; Liu and Zheng 2022; Grandvalet and Bengio 2004). Entropy minimization methods are founded on the clustering assumption, which posits that clusters should be compact within each class. Consistency regularization methods are founded on the smoothness assumption (Van Engelen and Hoos 2020), which posits that small perturbations should not produce the obvious deviations of corresponding outputs (Laine and Aila 2016). These methods often enforce that predictions are consistent across different views (e.g., different augmentations or perturbations) of the same data, thereby promoting the decision boundary to move towards low-density regions. Furthermore, FixMatch (Sohn et al. 2020) proposed a weak-strong consistency framework that combines entropy minimization and consistency regularization methods. Specifically, FixMatch encourages the model to produce consistent predictions across images with weak and strong augmentations. FixMatch provides a simple and effective approach for semi-supervised learning. Based on a similar idea, FlexMatch (Zhang et al. 2021) and FreeMatch (Wang et al. 2022b) suggest the adaptation of class-specific confidence thresholds to account for varying degrees of learning complexity. CoMatch (Li, Xiong, and Hoi 2021) and SimMatch (Zheng

et al. 2022) integrate contrastive learning objectives to leverage instance-level similarities.

Despite the success achieved in various tasks, all these methods can be summarized as instance perturbation-based methods. In this study, we propose a novel model perturbation and integrate it with instance perturbation to enhance the model's generalization ability towards unknown data, thereby improving model performance.

### Semi-supervised Medical Image Segmentation

Due to its prospect in real-world medical applications, SSL has also gained substantial attention from researchers in medical image analysis. For example, Luo et al. (2021) and Wang et al. (2022a) exploited the consistency among multiple correlated tasks, such as reconstruction, segmentation, and signed distance field prediction, to explore unlabeled data. Wu et al. (2022a) proposed the construction of multiple decoders, coupled with the application of a mutual consistency constraint among their predictions. This strategy is designed to encourage the model to generate invariant results. Luo et al. (2022) facilitates model learning by enforcing consistency between pyramid predictions with different scales. Wu et al. (2022b) utilized adversarial perturbations and contrastive learning approaches to improve pixel-level smoothness and inter-class separation in semi-supervised medical image segmentation. Lei et al. (2022) introduced ASE-Net in their work, leveraging adversarial consistency training combined with dynamic convolutions for semi-supervised medical image segmentation. More recently, Bai et al. (2023) introduced a highly effective data augmentation technique for combining labeled and unlabeled data, yielding impressive performance results. You et al. (2023a) adopted the concept of contrastive pre-training to address the scarcity and imbalance issues in medical imaging data, achieving outstanding performance.

Although these models have achieved good results in semi-supervised medical image segmentation, they still exhibit class overlap in samples from low-density regions, which adversely affects the segmentation performance. Therefore, in this work, we propose adversarial parameter perturbation, aiming to achieve low-density separation between classes, thus improving segmentation performance.

## Methodology

### Problem Setting

Given a limited labeled set $D_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ and a large amount of unlabeled set $D_u = \{u_j\}_{j=1}^{N_u}$, the goal of semi-supervised segmentation is to train a model parameterized by $\theta$ to map the input image $x$ to a segmentation mask $\hat{y}$ with satisfactory performance by utilizing a large amount of unlabeled data and limited labeled data. Here, $x_i$ and $y_i$ denote the limited labeled data and its label, respectively. $u_j$ denotes the image without annotations. $N_l$ and $N_u$ are the total numbers of limited labeled data and the large amount of unlabeled data.

### Preliminaries

In semi-supervised learning, where only a limited amount of labeled data is available, it is a standard practice to utilize this subset directly to guide the model's training. For a given labeled image $x$, the process begins by applying a data augmentation strategy to enhance the diversity of the labeled images. Subsequently, the model is used to generate a prediction $f(x;\theta)$. The supervised learning phase is then conducted using the supervised loss function:

$$\mathcal{L}_{\text{sup}} = \mathcal{H}(y, f(x;\theta)), \tag{1}$$

where $\mathcal{H}$ in this work is specifically defined as a combination of cross-entropy loss and dice loss.

Then, for unlabeled data, consistency regularization is widely applied by enforcing an invariance of predictions of input images under different perturbations and pushing the decision boundary to low-density regions, based on the assumptions that the perturbations should not change the output of the model. Several typical consistency regularization strategies are depicted in Fig. 2. In general, these methods can be divided into instance perturbation-based and model perturbation-based methods. Firstly, the instance perturbation-based consistency regularization methods usually add perturbations to each instance in input space or feature space, and further encourage the model to generate consistent predictions under perturbed. Representative methods include Virtual Adversarial Training (VAT (Miyato et al. 2018)) and Bidirectional Copy-Paste (BCP (Bai et al. 2023)). Formally, instance perturbation-based methods can be represented by:

$$\mathcal{L}_{\text{unsup}} = \mathcal{R}(f(u;\theta), f(u';\theta)) \tag{2}$$

where $u$ represents the unlabeled data and $u'$ represents the perturbed unlabeled data. And, $\mathcal{R}$ represents the consistency constraint.

Then, the model perturbation-based methods encourage the prediction to be consistent, which can be generated by different model parameters. For example, Luo et al. (2022) encourages consistency among pyramid outputs obtained via lateral connections from different layers, while Ouali, Hudelot, and Tami (2020) enforces consistency between multiple decoders. Formally, these methods can be represented by:

$$\mathcal{L}_{\text{unsup}} = \mathcal{R}(f(u;\theta), f(u;\theta')) \tag{3}$$

### Gradient-based Adversarial Perturbation (GAP)

The existing methods encourage consistency in predictions generated by different model parameters. However, they often rely on randomness (such as dropout) or prior knowledge (like multi-scale consistency). This reliance can lead to variability in performance, especially under varying conditions. To mitigate these issues, we propose Gradient-based Adversarial Perturbation (GAP), which aims to achieve better discrimination between classes in low-density regions.

Specifically, GAP first utilizes a vanilla consistency regularization method to calculate the consistency loss $\mathcal{L}_{\text{unsup}}$ to provide directional guidance for decision boundary movement. Then, GAP calculates the gradient of the model's parameters $g_1 = \nabla_\theta \mathcal{L}_{\text{unsup}}(\theta)$ based on $\mathcal{L}_{\text{unsup}}$, which would

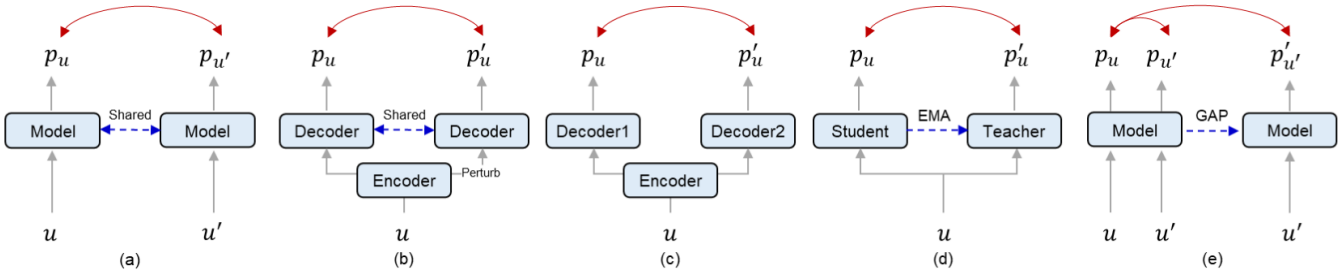


Figure 2: Comparison with the existing methods. (a)-(b)Instance perturbation-based consistency regularization from input space and feature space, respectively; (c)-(d) Model perturbation-based consistency regularization via multi-branch and teacher-student architecture; (e) Consistency regularization based on the proposed dual perturbation.

typically be updated using the gradient descent method. However, in this case, GAP employs gradient ascent to reverse update the model's parameters. Specifically, GAP add a perturbation $r = \epsilon \frac{g_1}{||g_1||}$ to the current model parameters such that $\theta^* = \theta + r$. By strategically manipulating the parameters in a counterintuitive manner, GAP effectively shifts the decision boundary in a direction contrary to its typical update trajectory. This operation expands the influence of the decision boundary, thereby augmenting its domain. Intuitively, this creates a stricter decision boundary for the model, especially for unlabeled instances. It is foreseeable that under a stricter decision boundary, subsequent consistency regularization will further push away instances that were initially near the boundary, distancing them from their original proximity.

Therefore, this method reduces the low-density region, which in turn strengthens the robustness of the decision boundary. As the boundary becomes more resilient, the classification reliability for low-confidence instances improves accordingly. In segmentation tasks, this enhanced method more clearly defines the target's boundary, leading to improved segmentation precision.

## Holistic Framework

Based on the proposed GAP, we propose a framework named GapMatch that focuses on combining the instance perturbation-based and model perturbation-based consistency regularization to enhance the robustness of the decision boundary and boost segmentation performance.

Firstly, for each unlabeled image, the weak augmentation (e.g. , random flip and crop) is applied to obtain weakly augmented unlabeled image $u^w$ and the model is used to predict on $u^w$ to generate the prediction

$$p_u^w = f\left(u^w; \theta\right). \tag{4}$$

Then, according to the predicted probabilities, the pseudo-label $\hat{y}_u$ is obtained by selecting the class with the highest probability as long as this probability surpasses a predefined confidence threshold $\tau$. Formally,

$$\hat{y}_u = \text{argmax}(p_u^w) \quad \text{if} \quad \max(p_u^w) \geq \tau. \tag{5}$$

Subsequently, a stronger augmentation (e.g. , color jitter) to the same unlabeled sample $u$, generating $u^s$, and obtaining the model's prediction result by

$$p_u^s = f\left(u^s; \theta\right). \tag{6}$$

The consistency regularization operator $\mathcal{R}$ is imposed on the model to enforce it generate the consistent prediction for $u^w$ and $u^s$:

$$\mathcal{L}_{\text{unsup}} = \mathcal{R}(\hat{y}_u, p_u^s), \tag{7}$$

where $\mathcal{R}$ is the combination of cross-entropy loss and dice loss in practice. Incorporating the vanilla consistency regularization is particularly advantageous because it establishes an avoidance zone around samples within the decision space, where decision boundary is not permitted to traverse these zones. This helps in progressively refining and pushing the decision boundary, allowing for a more nuanced and accurate classification of different types of samples.

Diverging from previous instance perturbation-based consistency regularization methods that directly update model parameters, our approach employs instance perturbation-based consistency as an initial condition, subsequently guiding the GAP process with adversarial perturbation direction. Since GAP is computed based on $\mathcal{L}_{\text{unsup}}$, it offers a more targeted perturbation method compared to others like dropout, effectively utilizing the unsupervised loss $\mathcal{L}_{\text{unsup}}$ to obtain perturbed model parameterized by $\theta^*$ and also establish a stricter decision boundary.

Then, the perturbed model with parameters $\theta^*$, is utilized again to predict $u^s$, resulting in:

$$p_u^{s\prime} = f(u^s; \theta^*). \tag{8}$$

To ensure robustness, a consistency regularization operator $\mathcal{R}$ is subsequently enforced on the outputs. This operation enforces the model to generate consistent predictions for $u^s$ under the influence of the perturbed parameters:

$$\mathcal{L}^*_{\text{unsup}} = \mathcal{R}(\hat{y}_u, p_u^{s\prime}). \tag{9}$$

In this way, we not only encourage that the model maintains prediction consistency across various instance perturbation-based perturbations, but also encourage that the model maintains prediction consistency when subjected to model perturbation-based perturbations. Instance perturbation-based consistency regularization pushes the decision boundary towards low-density areas, while consistency regularization based on model perturbation-based perturbation drives the sample away from the decision boundary, jointly achieving better class discrimination.

Algorithm 1: Optimization process of the proposed framework.

**Input**: Training set $D_l = \{(x_i, y_i)\}_{i=0}^{N_l}$, $D_u = \{u_i\}_{i=0}^{N_u}$; consistency loss $\mathcal{L}_{\text{unsup}}$; supervised loss $\mathcal{L}_{\text{sup}}$; total step $T$; balance coefficient $\alpha$; approximation scalar $r$; batch size $B$.
**Parameter**: Model parameters $\theta$.
**Output**: Optimized parameters $\hat{\theta}$.

Parameter initialization $\theta_0$.
**for** step $t = 1$ to $T$ **do**
  Fetch data $B_u = \{u_i\}_{i=0}^{B/2}$ and $B_l = \{(x_i, y_i)\}_{i=0}^{B/2}$
  Get augmented data $\{u_i^w\}_{i=0}^{B/2}, \{u_i^s\}_{i=0}^{B/2}$.
  Calculate the gradient $g_1 = \nabla_\theta \mathcal{L}_{\text{unsup}}(\theta)$ based on vanilla consistency regularization.
  Backup the parameter of the model.
  Add perturbation $\epsilon \frac{g_1}{||g_1||}$ on the current parameter $\theta$, which makes $\theta^* = \theta + \epsilon \frac{g_1}{||g_1||}$.
  Calculate the gradient $g_2 = \nabla_\theta \mathcal{L}_{\text{unsup}}(\theta)$ at $\theta = \theta^*$.
  Calculate gradient of unlabeled data $g_u = g_1 + \alpha g_2$.
  Recover the parameter of the model.
  Calculate gradient of labeled data $g_l = \nabla_\theta \mathcal{L}_{\text{sup}}(\theta)$.
  Calculate the final gradient $g = g_u + g_l$.
  Update parameter with the final gradient using the optimizer.
**end for**
**return** Final optimized parameters $\hat{\theta}$

Finally, the model learns from unlabeled data through two consistency losses, and the complete algorithmic process is depicted in Algorithm 1.

# Experiments

## Experimental Setup

The experiments are conducted on two public datasets: ACDC dataset (Bernard et al. 2018) and LA dataset (Xiong et al. 2021). The introduction and detailed information of the dataset will be described in the supplementary material. To evaluate the performance of methods, the overlap-based Dice coefficient (Dice) and surface-based discrepancy measure: Average Surface Distance(ASD) are reported. Dice is the larger the better but ASD is the smaller the better.

## Comparison with the State-of-the-arts

**Compared Methods.** To illustrate the effectiveness of the method, the proposed method was first compared with various competitive methods: including UNet and VNet using fully-supervised (FS) and limited supervised settings (LS); and semi-supervised learning methods: DAN (Zhang et al. 2017), DTC(Luo et al. 2021), DCT (Qiao et al. 2018), ICT (Verma et al. 2022), UAMT (Yu et al. 2019), SASSNet (Li, Zhang, and He 2020), CPS (Chen et al. 2021), GCL (Chaitanya et al. 2020), MC-Net (Wu et al. 2021), SS-Net (Wu et al. 2022b), ACTION (You et al. 2023b), BCP (Bai et al. 2023), ARCO (You et al. 2023a). The results of these methods were reported in the identical experimental setting in ARCO (You et al. 2023a).

**ACDC Results.** GapMatch was first compared with various competitive methods on the ACDC dataset. As shown in Table 1, GapMatch achieved superior segmentation performance across three anatomical sites. It is noteworthy that GapMatch trains models from scratch, yet it outperforms the previous SOTAs, i.e., ACTION and ARCO, which initialize parameters through pre-training, on the majority of metrics. This strongly demonstrates the effectiveness of our method.

Additionally, as the results listed in Table 1 show, we have drawn several observations as follows: 1) The performance of all the methods included in the table experienced a significant decline when the amount of available labeled data was reduced from 10% to 1%. However, GapMatch maintains the most robust performance among these methods, with the average performance decreasing from (90.1, 0.39) to (86.8, 0.81), a change of ($\downarrow$ 3.3, $\uparrow$ 0.42). 2) With only 1% of labeled data used for training, our method's results (86.8, 0.81) not only outperformed SS-Net's under the same conditions (63.4, 2.94) but also exceeded SS-Net's performance when it was trained with 10% of labeled data (86.8, 1.40). These observations demonstrate that GapMatch has established a better decision boundary for different classes, thus leading to high and robust performance.

**LA Results.** Compared with the fully-supervised methods, when using the same 1% of labeled data, GapMatch doubled the performance compared to VNet. Additionally, when using 10% of labeled data, GapMatch achieved a Dice score of 91.0, which is already close to the performance of the fully supervised VNet method (91.5). More encouragingly, GapMatch surpasses the fully supervised method in terms of ASD, providing compelling evidence of the efficacy of the proposed method in handling pixels at boundaries with inherent ambiguity. This result not only highlights the robustness of GapMatch but also emphasizes its potential as a reliable alternative to fully supervised methods, particularly when annotated data is scarce or expensive to obtain.

Additionally, the performance of GapMatch fully surpasses other semi-supervised methods for medical imaging. Specifically, the performance (88.3,1.77) achieved by GapMatch using only 5% labeled data is comparable to that of SS-Net using 10% labeled data (88.6, 1.90). Compared with some strong competitors (BCP, ARCO, ACTION), despite not employing complex augmentation strategies or pre-training, GapMatch still achieved superior performance. Moreover, under extremely limited data conditions, we have significantly outperformed these methods, achieving improvements of 3.0 and 1.25 in Dice score and ASD, respectively.

**Visual Results.** We present the predictive results of various semi-supervised medical image segmentation methods for a challenging case in Fig. 3. First, the comparison between the model predictions and the ground truth reveals that GapMatch is the only one that accurately predicts the morphology of the left atrium. Secondly, from the uncertainty map, it can be seen that GapMatch exhibits high confidence in predicting most regions of the left atrium, with only a small amount of high-entropy predictions occurring along the boundaries. In contrast, the DTC and SS-Net exhibit substantial uncertainty in non-target background regions, which

| Method | ACDC Dataset | | | LA Dataset | | |
|---|---|---|---|---|---|---|
| | 1% | 5% | 10% | 1% | 5% | 10% |
| FS | 91.5, 0.57 | 91.5, 0.57 | 91.5, 0.57 | 91.5, 1.51 | 91.5, 1.51 | 91.5, 1.51 |
| LS | 27.7, 32.6 | 56.6, 6.15 | 83.8, 2.84 | 40.0, 21.2 | 52.6, 9.87 | 82.7, 3.26 |
| DAN (Zhang et al. 2017) | 48.9, 17.5 | 56.4, 15.1 | 76.5, 3.01 | 38.5, 22.0 | 78.8, 6.53 | 80.2, 5.37 |
| DTC(Luo et al. 2021) | 51.7, 17.5 | 56.9, 7.59 | 84.3, 4.04 | 36.2, 11.7 | 83.6, 2.81 | 87.1, 2.23 |
| DCT (Qiao et al. 2018) | 49.7, 16.4 | 58.5, 10.8 | 78.1, 2.64 | 42.9, 19.1 | 80.1, 9.06 | 80.4, 9.18 |
| UAMT (Yu et al. 2019) | 36.9, 15.2 | 48.3, 9.14 | 81.8, 4.04 | 60.3, 11.3 | 82.3, 3.82 | 87.8, 2.12 |
| SASSNet (Li, Zhang, and He 2020) | 42.6, 24.8 | 57.8, 6.36 | 84.7, 1.83 | 51.5, 14.6 | 81.6, 3.58 | 87.5, 2.59 |
| CPS (Chen et al. 2021) | 51.5, 15.3 | 61.0, 2.92 | 78.8, 3.41 | 45.1, 22.0 | 79.7, 9.28 | 80.7, 5.16 |
| GCL (Chaitanya et al. 2020) | 59.7, 14.3 | 70.6, 2.24 | 87.0, 0.75 | 52.6, 12.8 | 75.5, 7.60 | 84.8, 4.22 |
| MC-Net (Wu et al. 2021) | 53.4, 17.1 | 62.8, 2.59 | 86.5, 1.89 | 44.3, 14.1 | 83.6, 2.70 | 87.6, 1.82 |
| SS-Net (Wu et al. 2022b) | 63.4, 2.94 | 65.8, 2.28 | 86.8, 1.40 | 43.4, 14.8 | 86.3, 2.31 | 88.6, 1.90 |
| ACTION (You et al. 2023b) | 81.0, 3.45 | 86.6, 1.20 | 87.2, 1.47 | 71.1, 6.23 | 86.6, 2.24 | 88.7, 2.10 |
| BCP (Bai et al. 2023) | 78.6, 1.73 | 87.6, 0.67 | 88.4, 1.17 | 76.8, 4.31 | 88.0, 2.15 | 89.6, 1.76 |
| ARCO (You et al. 2023a) | 85.5, 0.95 | 88.7, 0.84 | 89.4, 0.78 | 75.0, 4.06 | 87.8, **1.66** | 89.9, 1.47 |
| GapMatch (Ours) | **86.8**, **0.81** | **88.8**, **0.64** | **90.1**, **0.39** | **79.8**, **3.06** | **88.3**, 1.77 | **91.0**, **1.46** |

Table 1: Comparisons with state-of-the-art methods on the ACDC and LA datasets. The best and second-best metrics are shown in **bold** and underline, respectively. The segmentation performance is reported in terms of Dice (%) and ASD (voxel).

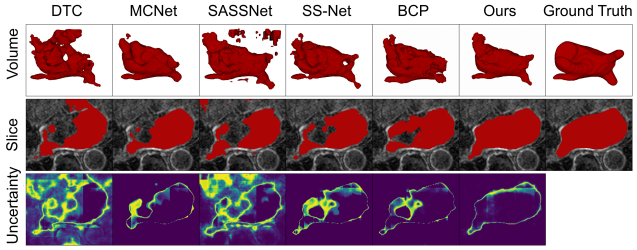


Figure 3: Visualizations of several semi-supervised segmentation methods with 10% labeled data and ground truth on LA dataset.

potentially indicates their limitations in distinguishing between foreground and background.

## Combination with Instance Perturbations

The proposed GAP is plug-and-play, hence various SSL methods can be integrated with it and benefit from it. To illustrate its compatibility and effectiveness, we have integrated GAP with several existing methods: entropy minimization (EM) (Vu et al. 2019), Pseudo Labeling, Interpolation Consistency Training (ICT) (Verma et al. 2022), MixUp, CutMix, and Virtual Adversarial Training (VAT) (Miyato et al. 2018). Among these, EM and Pseudo Labeling are entropy minimization-based methods, whereas the others fall under the category of consistency regularization methods. Specifically, when combined with GAP, we employ their respective methods to calculate the gradient direction required by GAP, followed by the application of perturbations. We further encourage the prediction to be aligned before and after perturbation.

As shown in Table 2, it can be seen that all methods have demonstrated performance improvements after incorporating the GAP. Notably, with 1% labeled data, EM, Pseudo Label, ICT, MixUp, CutMix, VAT increased Dice by 13.9%, 25.4%, 44.1%, 43.0%, 16.1%, 7.6%, respectively, compared to the baseline methods. Meanwhile, we observed that all methods exhibited enhanced stability across multiple experimental settings with GAP. The performance of Pseudo Label using 1% labeled data was nearly half of that using 10%, but the degree of performance degradation was significantly reduced after the addition of GAP. It is also worth mentioning that FixMatch even demonstrated performance comparable to that with 10% labeled data when only 1% was available.

| Method | 1% | 5% | 10% |
|---|---|---|---|
| EM | 23.7, 36.4 | 56.1, 5.52 | 84.9, 2.90 |
| + GAP | 37.6, 17.7 | 75.0, 0.82 | 88.1, 1.23 |
| Gain (△) | **↑13.9, ↓18.7** | **↑18.9, ↓4.7** | **↑3.2, ↓1.67** |
| Pseudo Label | 45.5, 6.97 | 64.4, 4.05 | 86.8, 1.66 |
| + GAP | 70.9, 1.21 | 84.8, 1.99 | 89.1, 0.62 |
| Gain (△) | **↑25.4, ↓5.76** | **↑20.4, ↓2.05** | **↑2.35, ↓1.04** |
| ICT | 31.7, 29.0 | 58.7, 4.71 | 84.7, 2.23 |
| + GAP | 75.8, 2.18 | 85.2, 1.10 | 88.1, 1.08 |
| Gain (△) | **↑44.1, ↓26.8** | **↑26.5, ↓3.61** | **↑3.4, ↓1.15** |
| MixUp | 33.4, 26.6 | 59.4, 2.66 | 84.4, 2.84 |
| + GAP | 76.4, 2.30 | 85.0, 1.76 | 88.0, 1.00 |
| Gain (△) | **↑43.0, ↓24.3** | **↑25.6, ↓0.9** | **↑3.6, ↓1.84** |
| CutMix | 42.8, 16.7 | 57.5, 7.86 | 84.5, 1.94 |
| + GAP | 58.9, 7.24 | 74.4, 1.32 | 87.4, 1.01 |
| Gain (△) | **↑16.1, ↓9.46** | **↑16.9, ↓6.54** | **↑2.9, ↓0.93** |
| VAT | 54.2, 2.99 | 70.3, 2.52 | 86.5, 1.12 |
| + GAP | 61.8, 3.10 | 75.5, 0.61 | 88.0, 0.91 |
| Gain (△) | **↑7.6, ↑0.11** | **↑5.2, ↓1.91** | **↑1.5, ↓0.21** |

Table 2: Ablation experiments on parameter perturbation in two commonly used semi-supervised methods on the ACDC dataset. The segmentation performance is reported in terms of Dice (%) and ASD (voxel).

| Method | 1% | 5% | 10% |
|---|---|---|---|
| LS | 27.7, 32.6 | 56.6, 6.15 | 83.8, 2.84 |
| MT | 27.5, 39.3 | 57.7, 11.4 | 82.9, 3.26 |
| URPC | 26.5, 14.4 | 50.3, 5.84 | 85.1, 1.39 |
| CCT | 25.3, 23.2 | 62.0, 3.25 | 87.6, 1.38 |
| ICT | 31.7, 29.0 | 58.7, 4.71 | 84.7, 2.23 |
| + GAP | 75.8, 2.18 | 85.2, 1.10 | 88.1, 1.08 |

Table 3: Comparisons with other model perturbations. The segmentation performance is reported in terms of Dice (%) and ASD (voxel).

## Comparison with Model Perturbations

The main idea of GAP is to keep the consistent prediction under different parameters, therefore we compare GAP with several model perturbations, including Mean Teacher (MT) (Tarvainen and Valpola 2017), Uncertainty Rectified Pyramid Consistency (URPC) (Luo et al. 2022), and Cross-Consistency Training (CCT) (Ouali, Hudelot, and Tami 2020). These methods regularize the learning of the model by enforcing consistency between the predictions of student and teacher models, consistency across multi-scale prediction results, consistency among predictions from multi-branch architectures, and consistency between the prediction results from models before and after feature perturbation, thereby enhancing the models' generalization prowess and robustness.

Due to the necessity of directional guidance in GAP, we adopt a naïve method, i.e., ICT, to regularize the training of unlabeled data and compute the adversarial direction for GAP. The detailed comparison across multiple experimental settings is shown in Table 3. Initially, when only limited labeled data is available, notably when just 1% labeled data is utilized, semi-supervised methods such as MT, URPC, and CCT do not demonstrate substantial performance enhancements compared to baselines trained exclusively on labeled data. This demonstrates that the perturbations introduced by these methods are insufficient to significantly improve the model's performance. Subsequently, ICT marginally outperforms baseline methods; however, with the integration of GAP, it witnesses a substantial boost in performance, notably achieving a 44.1% enhancement when only 1% labeled data is available. This is because GAP utilizes gradient information to compute the direction of perturbations, in contrast to strategies such as random dropout that rely on stochastic perturbations inherent to the model, thereby exhibiting greater specificity and enabling a more accurate regularization of model learning.

Moreover, Fig. 4 shows the visual comparison results of these model perturbations and GAP. When training with only 1% and 5% of the data, some methods partially or completely missed the area to be predicted, resulting in omissions in the prediction results. Upon increasing the data to 10%, all methods except for GAP introduced additional false positive predictions. In contrast, the GAP method consistently provided a relatively complete and precise segmentation of the foreground area, fully demonstrating the unique

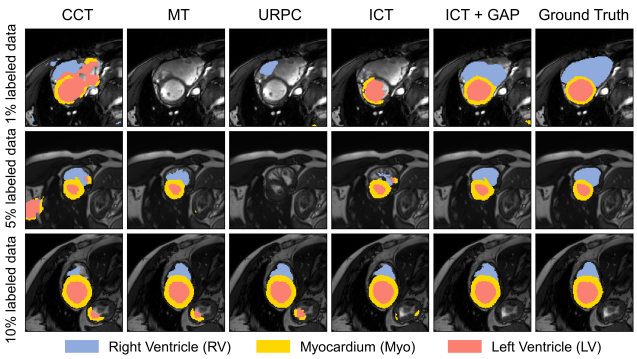


Figure 4: Visual comparison several of several model perturbations and GAP across multiple labeled ratios.

| Method | Train (s/iteration) | Inference(s/case) |
|---|---|---|
| UNet | 0.1679 | 0.6354 |
| MT | 0.1999 | 0.6363 |
| URPC | 0.2745 | 0.6322 |
| CCT | 0.5597 | 0.9728 |
| GapMatch | 0.3972 | 0.6303 |

Table 4: Training and Inference Times of Various Methods.

superiority of GAP in enhancing model performance across various labeled settings.

## Efficiency Analysis

We compared the training and inference speed of several model perturbation-based consistency regularization methods. As shown in Table 4, although all methods inevitably increase the training time, GapMatch has demonstrated its unique advantage in maintaining the inference speed. GapMatch takes 0.3972 seconds per iteration during the training phase, although this is an increase compared to other methods, it is within an acceptable range. Moreover, in contrast to methods that require altering the network architecture, such as CCT, GapMatch maintains its performance in the inference phase without compromising speed, taking 0.6303 seconds per case.

## Conclusion

In this study, we introduced an iterative dual perturbation-based consistency regularization method for semi-supervised medical image segmentation, which incorporates additional adversarial model perturbation with the commonly utilized instance perturbation. This dual-perturbation strategy encourages the model to maintain consistent predictions even after being perturbed at both the instance and model levels. Through this method, the model is encouraged to find decision boundary with greater class margins through model perturbation, so as to enhance the model's generalization on unseen data and achieve performance improvement. Extensive experiments conducted on two public datasets, along with qualitative and quantitative analyses, have demonstrated the effectiveness of our method.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation for Distinguished Young Scholar under Grant No.62025601 and in part by the National Natural Science Foundation Regional Innovation and Development Joint Fund under Grant No.U24A20341.

## References

Bai, Y.; Chen, D.; Li, Q.; Shen, W.; and Wang, Y. 2023. Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11514–11524.

Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; Yang, X.; Heng, P.-A.; Cetin, I.; Lekadir, K.; Camara, O.; Ballester, M. A. G.; et al. 2018. 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.

Chaitanya, K.; Erdil, E.; Karani, N.; and Konukoglu, E. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in neural information processing systems*, 33: 12546–12558.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2613–2622.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17.

Huang, W.; Zhang, L.; Wang, Z.; and Wang, L. 2024. Exploring Inherent Consistency for Semi-supervised Anatomical Structure Segmentation in Medical Imaging. *IEEE Transactions on Medical Imaging*.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Laine, S.; and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Lei, T.; Zhang, D.; Du, X.; Wang, X.; Wan, Y.; and Nandi, A. K. 2022. Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. *IEEE Transactions on Medical Imaging*.

Li, J.; Xiong, C.; and Hoi, S. C. 2021. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9475–9484.

Li, S.; Zhang, C.; and He, X. 2020. Shape-aware semi-supervised 3D semantic segmentation for medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, 552–561. Springer.

Liu, P.; and Zheng, G. 2022. Handling Imbalanced Data: Uncertainty-guided Virtual Adversarial Training with Batch Nuclear-norm Optimization for Semi-supervised Medical Image Classification. *IEEE Journal of Biomedical and Health Informatics*.

Luo, X.; Chen, J.; Song, T.; and Wang, G. 2021. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, 8801–8809.

Luo, X.; Wang, G.; Liao, W.; Chen, J.; Song, T.; Chen, Y.; Zhang, S.; Metaxas, D. N.; and Zhang, S. 2022. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80: 102517.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1979–1993.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12674–12684.

Qiao, S.; Shen, W.; Zhang, Z.; Wang, B.; and Yuille, A. 2018. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, 135–152.

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

Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J. N.; Wu, Z.; and Ding, X. 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63: 101693.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Van Engelen, J. E.; and Hoos, H. H. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2): 373–440.

Verma, V.; Kawaguchi, K.; Lamb, A.; Kannala, J.; Solin, A.; Bengio, Y.; and Lopez-Paz, D. 2022. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145: 90–106.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2517–2526.

Wang, K.; Zhan, B.; Zu, C.; Wu, X.; Zhou, J.; Zhou, L.; and Wang, Y. 2022a. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Medical Image Analysis*, 79: 102447.

Wang, P.; Peng, J.; Pedersoli, M.; Zhou, Y.; Zhang, C.; and Desrosiers, C. 2021a. Self-paced and self-consistent co-training for semi-supervised image segmentation. *Medical Image Analysis*, 73: 102146.

Wang, X.; Chen, H.; Xiang, H.; Lin, H.; Lin, X.; and Heng, P.-A. 2021b. Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. *Medical image analysis*, 70: 102010.

Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; et al. 2022b. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*.

Wang, Y.; Feng, Y.; Zhang, L.; Zhou, J. T.; Liu, Y.; Goh, R. S. M.; and Zhen, L. 2022c. Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images. *Medical Image Analysis*, 81: 102535.

Wu, Y.; Ge, Z.; Zhang, D.; Xu, M.; Zhang, L.; Xia, Y.; and Cai, J. 2022a. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81: 102530.

Wu, Y.; Wu, Z.; Wu, Q.; Ge, Z.; and Cai, J. 2022b. Exploring Smoothness and Class-Separation for Semi-supervised Medical Image Segmentation. *arXiv preprint arXiv:2203.01324*.

Wu, Y.; Xu, M.; Ge, Z.; Cai, J.; and Zhang, L. 2021. Semi-supervised left atrium segmentation with mutual consistency training. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, 297–306. Springer.

Xiong, Z.; Xia, Q.; Hu, Z.; Huang, N.; Bian, C.; Zheng, Y.; Vesal, S.; Ravikumar, N.; Maier, A.; Yang, X.; et al. 2021. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67: 101832.

You, C.; Dai, W.; Min, Y.; Liu, F.; Zhang, X.; Feng, C.; Clifton, D. A.; Zhou, S. K.; Staib, L. H.; and Duncan, J. S. 2023a. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *arXiv preprint arXiv:2302.01735*.

You, C.; Dai, W.; Min, Y.; Staib, L.; and Duncan, J. S. 2023b. Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. In *International Conference on Information Processing in Medical Imaging*, 641–653. Springer.

Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; and Heng, P.-A. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 605–613. Springer.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419.

Zhang, Y.; Yang, L.; Chen, J.; Fredericksen, M.; Hughes, D. P.; and Chen, D. Z. 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International conference on medical image computing and computer-assisted intervention*, 408–416. Springer.

Zheng, M.; You, S.; Huang, L.; Wang, F.; Qian, C.; and Xu, C. 2022. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14471–14481.