DiM: f-Divergence Minimization Guided Sharpness-Aware Optimization for Semi-supervised Medical Image Segmentation

Bingli Wang
Tsinghua University(SZ)

Houcheng Su
University of Macau

Nan Yin
Zayed University of Artificial Intelligence
United Arab Emirates
yinnan8911@gmail.com

Mengzhu Wang *
Hebei University of Technology
dreamkily@gmail.com

Li Shen *
Sun Yat-sen University
mathshenli@gmail.com

Abstract

As a technique to alleviate the pressure of data annotation, semi-supervised learning (SSL) has attracted widespread attention. In the specific domain of medical image segmentation, semi-supervised methods (SSMIS) have become a research hotspot due to their ability to reduce the need for large amounts of precisely annotated data. SSMIS focuses on enhancing the model's generalization performance by leveraging a small number of labeled samples and a large number of unlabeled samples. The latest sharpness-aware optimization (SAM) technique, which optimizes the model by reducing the sharpness of the loss function, has shown significant success in SSMIS. However, SAM and its variants may not fully account for the distribution differences between different datasets. To address this issue, we propose a sharpness-aware optimization method based on f-divergence minimization (DiM) for semi-supervised medical image segmentation. This method enhances the model's stability by fine-tuning the sensitivity of model parameters and improves the model's adaptability to different datasets through the introduction of f-divergence. By reducing f-divergence, the DiM method not only improves the performance balance between the source and target datasets but also prevents performance degradation due to overfitting on the source dataset.

1. Introduction

Medical Image Segmentation (MIS)[3, 14, 33] plays a crucial role in assisting computers with disease diagnosis and treatment research by helping identify key organs or lesions in abnormal images. Recently, numerous supervised learning-based encoder-decoder network architectures have made significant advancements in medical image segmentation, such as U-Net[24], U-Net++[46], and H-DenseUNet[13]. However, the success of these technologies largely relies on large-scale, pixel-level annotated data. In practice, annotating medical images is not only costly but also challenging due to issues such as low contrast and noise, making it difficult to clearly display images. Moreover, medical images require more specialized knowledge compared to natural images, which makes constructing a large-scale, accurately annotated medical image database nearly an impossible task. In contrast, semi-supervised learning [23, 35] offers a new solution to the problem of insufficient data supervision in weakly supervised learning [47]. It primarily utilizes a small amount of labeled data and a large amount of unlabeled data for joint training. Clearly, semi-supervised learning is significantly more suitable for medical image segmentation and adapting to real-world clinical scenarios than traditional supervised learning methods.

Due to the easy availability of unlabeled data, doctors may not have the time to verify its distribution when faced with massive amounts of data. This "domain shift" [7, 25] issue can lead to significant performance degradation in models, and it is a critical concern when developing semi-supervised medical image segmentation (SSMIS) [21] models. In fact, we should allow unlabeled data to come from one or more different distributions. However, existing unsupervised domain adaptation (UDA) [11, 43, 44] methods do not directly address this issue because they rely on large amounts of labeled source domain data, which is exactly what SSMIS aims to resolve.

Recent studies, such as Sharpness-Aware Minimization (SAM)[8], enhance model generalization performance by reducing the sharpness of the loss function. Here, $\mathcal{L}$ repre-

---

*Corresponding Author

sents the loss function to be minimized, and θ represents the parameters of the neural network. SAM first computes a weight perturbation ϵ that maximizes the empirical risk $\mathcal{L}(\theta)$, and then minimizes the loss of the perturbed network. In short, SAM aims to reduce the maximum loss near the model parameters θ. Due to the complexity of this minimization-maximization optimization problem, SAM approximates $\mathcal{L}$ with a surrogate loss function $\mathcal{L}_p(\theta)$ for minimization. However, it is important to note that minimizing $\mathcal{L}_p(\theta)$ does not guarantee reaching the flat minimum region for SSMIS [48]. KL divergence [32] has demonstrated strong performance in SSMIS. The application of KL divergence in SSMIS primarily improves model training efficiency and accuracy by measuring the differences between different probability distributions. This is particularly useful when dealing with limited labeled data and a large amount of unlabeled data, as it helps guide the model to learn more useful information in a semi-supervised setting. For example, MMLBF [6] propose a region-based multi-phase level set method based on KL divergence. Lu et al.[18] estimate uncertainty by calculating the Kullback-Leibler divergence between the predictions of the student and teacher models, and directly use this uncertainty to correct the learning of noisy pseudo-labels, rather than setting a fixed threshold to filter pseudo-labels. SwinMM[37] includes a masked multi-view encoder and a novel proxy task based on mutual learning, which contributes to effective self-supervised pretraining.

However, all of these methods are considered from the perspective of KL divergence, which is highly sensitive to probability values close to zero in the target distribution, often leading to an infinite divergence. In contrast, f-divergence can reduce this sensitivity by selecting an appropriate function, making it more stable and robust, especially when dealing with sparse or extreme distributions. In SSMIS, SAM emphasizes achieving stability by controlling the sensitivity of model parameters, while the introduction of f-divergence helps further regulate the model's adaptability across different domains. By minimizing f-divergence, SAM can enhance the balanced performance of the model across both the source and target domains, while avoiding performance degradation due to overfitting the source domain. In this work, to overcome the limitation of SAM and explore the full potential of f-divergence, we present a novel method f-divergence minimization guided sharpness-aware optimization for semi-supervised medical image segmentation (DiM). By consider the f-divergence and sharpness-aware minimization, which can still be effectively computed even when the support sets of the distributions are different. Our main contributions can be summarized as follows:

- we analyze the limitations of SAM-like methods and propose f-divergence to ensure the model convergence to a

flat region with a small loss.
- To the best of our knowledge, this is the first work to apply f-divergence constraints to SAM paradigm.
- we demonstrate the superior performance of DiM to state-of-the-arts on three SSMIS benchmarks.

2. Related Work

2.1. Semi-supervised Medical Image Segmentation

Due to the complexity of medical images, extensive manual annotation by experts is both challenging and costly [49]. To address this, semi-supervised medical image segmentation approaches have emerged as effective solutions that leverage limited labeled data [3]. Luo et al. [19] proposed a dual-task consistency-based semi-supervised framework to simultaneously predict per-pixel segmentation maps and geometrically-aware level set representations, introducing a dual-task consistency regularization to enhance performance. Wu et al. [39] presented MC-Net++, which employs a shared encoder and multiple distinct decoders and introduced a new mutual consistency constraint. This approach statistically identifies uncertain regions, particularly hard regions within unlabeled data. Luo et al. [20] incorporated a cross-teaching approach between CNNs and Transformers, resulting in a simple yet efficient semi-supervised learning framework. Miao et al. [23] highlighted the importance of model independence between networks or branches in semi-supervised medical segmentation (SSMS). Ma et al. [22] identified the issue of performance degradation in semi-supervised medical image segmentation due to shared domain distributions, proposing Mixed-domain Semi-supervised Medical Image Segmentation (MIDSS). They emphasized that generating reliable pseudo-labels for unlabeled data is crucial in domain shifts in labeled data. Nevertheless, due to the inherent complexity of medical images, these models often exhibit limited generalization ability and convergence instability, which remains a challenging problem.

2.2. Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) [9, 15, 34] aims to adapt models from a labeled source domain to an unlabeled target domain by minimizing the domain shift. This alignment of feature distributions enables knowledge transfer from source to target, enhancing classification performance [10, 38]. Many UDA approaches use a domain classifier to distinguish source from target features, while the feature extractor learns to match feature distributions [16, 30]. UDA is widely used in tasks like image classification [15], semantic segmentation [28], and object detection [27]. Semi-supervised domain adaptation further incorporates a small amount of labeled target data to improve transfer [26].

Table 1. Various commonly used f-divergences with their derivatives and second derivatives.

f-divergence	$f(x)$	$f'(x)$	$f''(x)$
Reverse KL	$x\log x$	$\log x + 1$	$\frac{1}{x}$
Forward KL	$-\log x$	$-\frac{1}{x}$	$\frac{1}{x^2}$
Jeffrey	$(x-1)\log x$	$\log x + 1 - \frac{1}{x}$	$\frac{1}{x} + \frac{1}{x^2}$
Jensen-Shannon	$-\frac{x+1}{2}\log\frac{x+1}{2} + \frac{x}{2}\log x$	$\frac{1}{2}\log\frac{2x}{x+1}$	$\frac{1}{2x(x+1)}$
Pearson	$\frac{(1-x)^2}{x}$	$1 - \frac{1}{x^2}$	$\frac{2}{x^3}$

2.3. Sharpness-Aware Minimization (SAM)

Foret et al. [8] observed that solely minimizing training loss can lead to suboptimal model quality. They proposed Sharpness-Aware Minimization (SAM), which seeks parameters in neighborhoods of uniformly low loss, resulting in a Min-Max optimization problem suitable for gradient descent. Andriushchenko et al. [1] provided theoretical insights on SAM's implicit bias in diagonal linear networks and empirically examined its behavior in nonlinear networks. Zhou et al. [45] addressed SAM's limitation in handling class imbalance, particularly overfitting to tail classes, by introducing Imbalanced-SAM (ImbSAM), a class-aware smoothing approach effective in long-tailed classification and semi-supervised anomaly detection tasks. Wang et al. [36] introduced a model integrating MedSAM with an uncertainty-aware loss function and SharpMin optimizer, enhancing segmentation accuracy and robustness. However, a tailored solution for semi-supervised medical image segmentation remains absent.

3. Methods

3.1. Aligning Features via f-Divergence

Semi-supervised medical image segmentation is challenging due to the scarcity of labeled data and the high-dimensional complexity of medical images. In this setting, models must leverage both labeled and unlabeled data to learn precise segmentation boundaries. However, limited labeled data can lead to feature drift, where the representations learned from unlabeled data deviate from those based on labeled data. This misalignment between labeled and unlabeled feature distributions reduces segmentation accuracy and limits generalization on unseen data.

To address this, we utilize f-divergence to align the high-dimensional logits from labeled and unlabeled data, constraining the features of unlabeled data based on a limited set of labeled data. Let p_{label} and p_{unlabel} denote the distributions of logits for the labeled and unlabeled data over a discrete set $\mathcal{X}$. Our goal is to guide p_{unlabel} by minimizing the f-divergence between these distributions in high-dimensional space.

This alignment mitigates feature drift, improving gener-

alization in the target domain. Additionally, f-divergence operates effectively in high-dimensional spaces, making it well-suited for capturing subtle but critical variations in medical image features. By aligning feature distributions, f-divergence fosters a stable and consistent feature representation, enhancing segmentation accuracy in semi-supervised conditions.

For a convex function $f(x) : \mathbb{R}^+ \to \mathbb{R}$ with $f(1) = 0$, the f-divergence $D_f(p_{\text{label}}\|p_{\text{unlabel}})$ is defined as:

$$
\begin{aligned}
D_f(p_{\text{label}}\|p_{\text{unlabel}}) = \mathbb{E}_{x\sim p_{\text{unlabel}}}&\left[f\left(\frac{p_{\text{label}}(x)}{p_{\text{unlabel}}(x)}\right)\right] \\
&+ f'(\infty)p_{\text{label}}(p_{\text{unlabel}} = 0),
\end{aligned}
\tag{1}
$$

where $f'(\infty) = \lim_{t\to 0} tf\left(\frac{1}{t}\right)$. The second term represents the contribution of points x in the support of p_{label} where $p_{\text{unlabel}}(x) = 0$, which accounts for cases where the labeled and unlabeled data distributions do not overlap.

To align p_{label} and p_{unlabel}, we define the alignment loss:

$$
\begin{aligned}
\mathcal{L}_{\text{align}} &= D_f(p_{\text{label}}\|p_{\text{unlabel}}) \\
&= \mathbb{E}_{x\sim p_{\text{unlabel}}}\left[f\left(\frac{p_{\text{label}}(x)}{p_{\text{unlabel}}(x)}\right)\right]
\end{aligned}
\tag{2}
$$

Minimizing $\mathcal{L}_{\text{align}}$ encourages the logits of unlabeled data to approximate those of labeled data, enhancing feature consistency for semi-supervised learning.

To quantify this alignment, we utilize specific f-divergence variants frequently used in machine learning, including Jeffrey divergence, Jensen-Shannon divergence, and Pearson divergence. Each variant has a unique form of $f(x)$, $f'(x)$, and $f''(x)$, as shown in Table 1, enabling flexible divergence calculations between p_{label} and p_{unlabel}. The f-divergences are computed via Monte Carlo estimation based on samples from p_{unlabel}, applying the respective values in Table 1 to evaluate $\mathcal{L}_{\text{align}}$ and facilitate backpropagation during training.

3.2. Sharpness-Aware Entropy Minimization

Semi-supervised medical image segmentation leverages both labeled and unlabeled data for accurate boundary detection. However, limited labeled data and distribution

shifts often lead to feature inconsistency and unreliable predictions, especially on unseen test samples, necessitating methods that improve model robustness to distributional variations. Sharpness-aware minimization (SAM) enhances model generalization by optimizing within low-loss neighborhoods, stabilizing performance under distribution shifts. However, directly filtering unreliable test samples using gradient norms is challenging due to variations in scale across models and shifts.

Directly using gradient norms to filter out unreliable test samples is challenging due to variability in scale across models and types of distribution shifts. Instead, we leverage entropy as a proxy for gradient magnitude, selecting samples with low entropy values to focus adaptation on confident predictions. Given an entropy function $E(\boldsymbol{x}; \boldsymbol{\theta})$ for a sample $\boldsymbol{x}$ with model parameters $\boldsymbol{\theta}$, we define the selective entropy minimization as:

$$\min_{\boldsymbol{\theta}} S(\boldsymbol{x})E(\boldsymbol{x}; \boldsymbol{\theta}), \quad S(\boldsymbol{x}) \triangleq \mathbb{I}_{\{E(\boldsymbol{x}; \boldsymbol{\theta}) < E_0\}}(\boldsymbol{x}) \quad (3)$$

where $S(\boldsymbol{x})$ is an indicator function that activates when the entropy $E(\boldsymbol{x}; \boldsymbol{\theta})$ is below a pre-defined threshold E_0. This approach ensures that only samples with low entropy (i.e., high confidence) contribute to the training, effectively filtering out unreliable samples that might otherwise induce large gradients.

For further stability, we aim to guide the model towards flatter regions of the entropy loss landscape, which reduces sensitivity to noisy gradients. We define a sharpness-aware entropy objective, $E^{\text{SA}}(\boldsymbol{x}; \boldsymbol{\theta})$, that measures the maximum entropy within a perturbation neighborhood around the current parameters:

$$\min_{\boldsymbol{\theta}} E^{\text{SA}}(\boldsymbol{x}; \boldsymbol{\theta}), \quad E^{\text{SA}}(\boldsymbol{x}; \boldsymbol{\theta}) \triangleq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} E(\boldsymbol{x}; \boldsymbol{\theta} + \boldsymbol{\epsilon}) \quad (4)$$

where $\boldsymbol{\epsilon}$ is a perturbation vector constrained within a Euclidean ball of radius ρ. This inner maximization encourages the model to be robust against perturbations, promoting a flat minimum for the entropy loss. Following the SAM approach, we approximate $\boldsymbol{\epsilon}^*(\boldsymbol{\theta})$ by:

$$\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}) = \rho \, \text{sign}\left(\nabla_{\boldsymbol{\theta}} E(\boldsymbol{x}; \boldsymbol{\theta})\right) \frac{|\nabla_{\boldsymbol{\theta}} E(\boldsymbol{x}; \boldsymbol{\theta})|}{\|\nabla_{\boldsymbol{\theta}} E(\boldsymbol{x}; \boldsymbol{\theta})\|_2} \quad (5)$$

Substituting $\hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})$ back into the objective, we obtain an approximation for the gradient that encourages flat minima:

$$\nabla_{\boldsymbol{\theta}} E^{\text{SA}}(\boldsymbol{x}; \boldsymbol{\theta}) \approx \nabla_{\boldsymbol{\theta}} E(\boldsymbol{x}; \boldsymbol{\theta})\Big|_{\boldsymbol{\theta} + \hat{\boldsymbol{\epsilon}}(\boldsymbol{\theta})} \quad (6)$$

Our final objective for Reliable Sharpness-Aware Entropy Minimization combines selective entropy minimization and sharpness-aware optimization:

$$\min_{\tilde{\boldsymbol{\theta}}} S(\boldsymbol{x})E^{\text{SA}}(\boldsymbol{x}; \boldsymbol{\theta}) \quad (7)$$

In this study, we introduce Sharpness-Aware Entropy Minimization(SAEM), an approach that combines entropy minimization with sharpness-aware training to achieve adaptive entropy reduction, enhancing model stability under challenging conditions. Here, $S(\boldsymbol{x})$ and $E^{\text{SA}}(\boldsymbol{x}; \boldsymbol{\theta})$ represent entropy measures as defined in Equations (3) and (4), respectively. The learnable parameters designated for adaptation are denoted as $\tilde{\boldsymbol{\theta}} \subset \boldsymbol{\theta}$.

In essence, SAEM offers a robust framework by integrating entropy filtering with sharpness-aware training, yielding adaptive entropy reduction while ensuring model resilience, particularly under demanding conditions.

3.3. Loss Function

The overall loss $\mathcal{L}_{\text{total}}$ is composed of the following components:

1. **Supervised Loss ($\mathcal{L}_s$)**: Applied to labeled data to guide predictions with ground truth, combining cross-entropy and dice losses for accurate segmentation.

2. **Intermediate Losses ($\mathcal{L}_{in}$ and $\mathcal{L}_{out}$)**: Defined for intermediate samples u_{in}^s and u_{out}^s, each using weighted cross-entropy ($\mathcal{L}_{ce}$) and dice loss ($\mathcal{L}_{dice}$) to enforce consistency between pseudo labels and model predictions.

$$\mathcal{L}_{in} = \mathcal{L}_{ce}(\hat{p}_{in}, p_{in}^s, w_{in}) + \mathcal{L}_{dice}(\hat{p}_{in}, p_{in}^s, w_{in}) \quad (8)$$

$$\begin{aligned} \mathcal{L}_{out} = &\mathcal{L}_{ce}(\hat{p}_{out}, p_{out}^s, w_{out}) \\ &+ \mathcal{L}_{dice}(\hat{p}_{out}, p_{out}^s, w_{out}) \end{aligned} \quad (9)$$

where w_{in} and w_{out} are pixel-wise weights set by a confidence threshold to filter unreliable pseudo labels.

Each component plays a crucial role in enforcing robust supervision on both labeled and unlabeled data, supporting reliable predictions across domains. The overall loss $\mathcal{L}_{\text{total}}$ is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_s + \lambda \left(\mathcal{L}_{in} + \mathcal{L}_{out} + \lambda \mathcal{L}_{\text{sym}}\right) + \mathcal{L}_{\text{align}} \quad (10)$$

where λ is a time-dependent coefficient that scales unsupervised components as training progresses, defined by:

$$\lambda(t) = e^{-5(1 - t/t_{\text{total}})}. \quad (11)$$

4. Experiments

4.1. Experiment Datasets

Fundus dataset consists of retinal fundus images gathered from four medical centers, mainly intended for tasks involving the segmentation of the optic cup and disc. Each image has been cropped to create a region of interest within an 800×800 bounding box. We then resize and randomly crop these images to a size of 256×256.

Prostate dataset includes prostate T2-weighted MRI data, complete with segmentation masks, sourced from six different locations across three public datasets. We randomly

divide the dataset into training and testing sets at a ratio of 4:1, resizing and randomly cropping each 2D slice to 384 × 384. Labeled samples are chosen from consecutive slices within individual cases, ensuring there is at most one case overlap and no overlap of slices with unlabeled samples.

4.2. Comparison Methods and Settings

Our method is implemented in PyTorch and utilizes an NVIDIA GeForce RTX 3090 GPU. We establish default experimental parameters for training. Optimization is performed using the SAM optimizer, with a base optimizer of Stochastic Gradient Descent (SGD) set to a momentum of 0.9, a weight decay of 0.0001, and an initial learning rate of 0.03. The batch size is set to 8, comprising 4 labeled and 4 unlabeled samples. We conduct a total of 30,000 iterations for the Fundus dataset and 60,000 iterations for the Prostate dataset.

During testing, the final segmentation results are generated by the student model. Our approach is benchmarked against several state-of-the-art (SOTA) methods, including supervised techniques such as UA-MT [42], FixMatch [31], CPS [5], CoraNet [29], SS-Net [40], BCP [2], CauSSL [23], and MiDSS [21], as well as domain-unsupervised adaptation methods like FDA [41], SIFA [4], and UDA-VAE++ [17].

In each experiment, a limited amount of data from a designated domain (e.g., Domain 1 in Tab. 2) is labeled, while the remaining data are treated as unlabeled. For the upper-bound comparison, we utilized the f-divergence, specifically employing the Jensen-Shannon divergence to calculate the distance between logits, and used the most naive SAM optimizer in our experiments. For the upper bound, we followed the results of the MiDSS paper, which applied UCP within the FixMatch framework, utilizing all available training data from a specific domain as labeled data, providing the model with comprehensive source domain information.

Our evaluation metrics include the Dice coefficient (DC), Jaccard coefficient (JC), 95% Hausdorff Distance (HD), and Average Surface Distance (ASD). Except for SIFA, which incorporates ResNet blocks [12] for its generator and decoder, all methods employ the U-Net backbone [24].

4.3. Comparison with State-of-the-Art Methods

Results on Fundus dataset. With only 20 labeled samples, DiM achieves superior performance across all domains in the optic cup/disc segmentation task, as illustrated in Table 2. DiM consistently outperforms competing methods in all metrics, which achieves the highest average DC and JC scores while maintaining the lowest HD and ASD, highlighting its robustness and precision in segmenting dual objects with overlapping regions. These results suggest that DiM effectively mitigates issues faced by other semi-

supervised and unsupervised domain adaptation methods, such as error accumulation and limited knowledge transfer, ensuring both high accuracy and generalizability across multiple domains.

Results on Prostate dataset. As shown in Table 3, DiM achieves outstanding performance across all metrics on Prostate dataset. It is also noteworthy that DiM achieves the highest DC and JC averages while maintaining the lowest HD and ASD scores, suggesting higher segmentation accuracy and boundary precision. The inclusion of SAM likely contributes to improved generalization by mitigating sharp minima, while f-divergence loss enhances alignment of the predicted and true distributions, reducing segmentation errors. These results underscore the robustness and effectiveness of our method.

4.4. Ablation Study

Firstly, We conduct ablation studies to show the impact of each component in DiM . Sencondly, our objective is to assess whether varying types of f-divergences lead to notable performance differences in the model and to identify the optimal form for superior outcomes.

The effectiveness of SAM. As demonstrated in Table 4, incorporating the Sharpness-Aware Minimization (SAM) optimizer (as seen in Method #2 and #4 compared to #1.) enhances model performance on the Optic Cup/Disc segmentation task. SAM effectively reduces the model's loss, increasing robustness to minor data distribution shifts and enabling more efficient capture of inter-sample similarity. Consequently, SAM improves the DC and JC scores while reducing the HD and ASD. These results indicate that SAM not only strengthens the model's generalization ability but also enhances segmentation accuracy and boundary precision.

The effectiveness of f-Divergence. The incorporation of f-divergence (as seen in Method #3 and #4 compared to #1.) contributes to notable improvements in model performance, particularly through a marked reduction in ASD and HD metrics, as shown in Table 4. By quantifying the discrepancy between probability distributions of labeled and unlabeled data, f-divergence enhances the model's capacity to represent features within the unlabeled dataset. Experimental results demonstrate that introducing f-divergence allows the model to more precisely capture the boundaries of structurally similar regions. This results in further gains in DC and JC metrics, along with substantial reductions in HD and ASD, thereby indicating improved accuracy in segmentation, especially along edge details.

f-**Divergence strategies.** Based on the experimental results shown in the Table 5, different f-divergence strategies demonstrate varying degrees of effectiveness for optic cup and disc segmentation across four domains in the Fundus dataset. Generally, the JS divergence and Jeffrey diver-

Task			Optic Cup / Disc Segmentation							
Method		#L	DC ↑				DC ↑	JC ↑	HD ↓	ASD ↓
			Domain 1	Domain 2	Domain 3	Domain 4	Avg.	Avg.	Avg.	Avg.
U-Net		20	59.54 / 73.89	71.28 / 74.23	50.87 / 64.29	35.61 / 63.30	61.63	52.65	48.28	28.86
UA-MT	MICCAI'19	20	59.35 / 78.46	63.08 / 74.45	35.24 / 47.73	36.18 / 55.43	56.24	47.00	48.64	31.35
FDA	CVPR'20	20	76.99 / 89.94	77.69 / 89.63	78.27 / 90.96	64.52 / 74.29	80.29	71.05	16.23	8.44
SIFA	TMI'20	20	50.67 / 75.30	64.44 / 80.69	61.67 / 83.77	55.07 / 70.67	67.78	54.77	20.16	10.93
FixMatch	NeurIPS'20	20	81.18 / 91.29	72.04 / 87.60	80.41 / 92.95	74.58 / 87.07	83.39	73.48	11.77	5.60
CPS	CVPR'21	20	64.53 / 86.25	70.26 / 86.97	42.92 / 54.94	36.98 / 46.70	61.19	52.69	34.44	26.79
CoraNet	TMI'21	20	61.64 / 87.32	65.56 / 87.05	66.12 / 83.54	49.01 / 77.73	72.25	60.50	20.52	10.44
UDA-VAE++	CVPR'22	20	55.01 / 80.76	68.87 / 85.94	63.23 / 84.92	68.42 / 80.89	73.51	61.40	17.60	9.86
SS-Net	MICCAI'22	20	59.42 / 78.15	67.32 / 85.05	45.69 / 69.91	38.76 / 61.13	63.18	53.49	44.90	25.73
BCP	CVPR'23	20	71.65 / 91.10	77.19 / **92.00**	72.63 / 90.77	77.67 / 91.42	83.05	73.66	11.05	5.80
CauSSL	ICCV'23	20	63.38 / 80.60	67.52 / 80.72	49.53 / 63.88	39.43 / 49.43	61.81	51.80	41.25	23.94
MiDSS	CVPR'24	20	<u>83.39</u> / **92.96**	73.12 / 88.88	<u>83.50</u> / 92.97	78.63 / **93.38**	<u>85.85</u>	<u>76.95</u>	<u>9.06</u>	<u>4.40</u>
DiM		20	**84.68** / 92.91	**78.16** / 90.49	**84.82** / 93.36	**81.63** / 92.18	**87.28**	**78.53**	**7.83**	**3.82**
Upper bound		*	85.53 / 93.41	80.55 / 90.90	85.44 / 93.04	85.61 / 93.21	88.46	80.35	7.41	3.70

Table 2. Comparison of methods on the Fundus dataset. #L indicates the number of labeled samples. In the "Upper bound" row, * denotes using all training samples in a domain as labeled data. An upward arrow (↑) signifies that higher values indicate better performance, while a downward arrow (↓) indicates the opposite. The best results are bolded, with the second-best underlined.

Task			Prostate Segmentation									
Method		#L	DC ↑						DC ↑	JC ↑	HD ↓	ASD ↓
			RUNMC	BMC	HCRUDB	UCL	BIDMC	HK	Avg.	Avg.	Avg.	Avg.
U-Net		40	31.11	35.07	20.04	38.18	19.41	26.62	28.41	23.24	95.11	65.84
UA-MT	MICCAI'19	40	29.44	4.68	12.49	39.42	17.94	18.22	20.37	14.88	112.07	77.58
FDA	CVPR'20	40	47.44	35.37	24.54	61.01	28.19	40.51	39.51	32.17	76.67	47.87
SIFA	TMI'20	40	72.67	70.37	64.08	73.49	71.62	65.16	69.57	56.78	29.43	13.03
FixMatch	NeurIPS'20	40	83.58	69.17	73.63	79.21	56.07	84.78	74.41	65.96	24.18	14.09
CPS	CVPR'21	40	29.83	9.21	11.84	43.84	13.51	14.56	20.47	15.12	115.96	78.51
CoraNet	TMI'21	40	69.43	31.16	16.29	69.33	24.66	22.16	38.84	31.48	67.91	44.98
UDA-VAE++	CVPR'22	40	68.73	69.36	65.49	67.19	63.29	65.15	66.54	52.80	34.20	15.48
SS-Net	MICCAI'22	40	29.10	13.49	14.20	51.96	23.83	13.23	24.30	18.74	109.54	71.13
BCP	CVPR'23	40	70.15	71.97	46.15	58.93	74.21	67.47	64.81	55.17	52.60	27.22
CauSSL	ICCV'23	40	24.10	27.46	16.94	27.23	15.28	14.56	20.93	15.48	114.62	73.30
MiDSS	CVPR'24	40	<u>87.94</u>	<u>85.30</u>	<u>77.74</u>	<u>86.29</u>	**88.54**	<u>86.43</u>	<u>85.37</u>	<u>77.30</u>	<u>13.44</u>	<u>6.18</u>
DiM		40	**88.43**	**85.67**	**87.56**	**87.27**	<u>88.23</u>	**87.55**	**87.45**	**79.52**	**10.57**	**4.35**
Upper bound		*	88.52	88.61	85.71	88.61	88.98	89.49	88.32	80.71	10.05	4.12

Table 3. Comparison of different methods on Prostate dataset.

Task				Optic Cup / Disc Segmentation			
Method	Baseline	SAM	f-Divergence	DC ↑	JC ↑	HD ↓	ASD ↓
#1	✓			88.27	80.02	7.19	3.48
#2	✓	✓		88.33	80.22	7.10	3.52
#3	✓		✓	88.32	80.14	7.22	3.47
#4	✓	✓	✓	**88.80**	**80.74**	**6.81**	**3.27**

Table 4. Ablation experiments across domain 1 on Fundus dataset.

gence strategies perform well, often yielding higher DC and JC while reducing HD and ASD. Specifically, JS divergence tends to provide more consistent results in edge cases, as evidenced by lower HD and ASD values, while also achieving competitive segmentation accuracy. Pearson divergence, on the other hand, exhibits a balanced performance, particularly excelling as the second-best in some metrics.
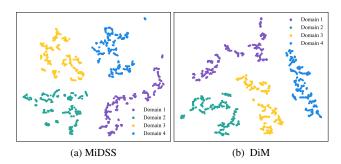
4.5. Visualization Analysis

(a) MiDSS (b) DiM

Figure 1. A T-sne visualization analysis was performed on the Fundus dataset experiment.

T-SNE visualization. We adopt the T-SNE visualization method , which graphically represents the learning rep-

Task		Optic Cup / Disc Segmentation							
Type	Domain	DC ↑		JC ↑		HD ↓		ASD ↓	
		Cup	Disc	Cup	Disc	Cup	Disc	Cup	Disc
baseline	Domain 1	83.38	**93.15**	72.68	**87.36**	8.28	<u>6.10</u>	4.01	**2.95**
JS		**84.68**	92.91	**74.51**	86.96	**7.63**	**5.98**	**3.56**	<u>2.97</u>
Jeffrey		82.82	92.57	72.07	86.45	8.20	6.21	4.13	3.20
Pearson		<u>83.54</u>	<u>92.94</u>	<u>73.15</u>	<u>87.07</u>	<u>7.91</u>	6.23	<u>3.94</u>	3.09
baseline	Domain 2	73.11	88.88	59.76	80.78	12.89	13.69	6.56	6.11
JS		78.10	89.32	<u>65.54</u>	81.28	10.71	13.72	5.29	6.21
Jeffrey		<u>78.16</u>	**90.49**	65.44	**82.99**	**9.62**	**8.24**	**4.83**	**4.37**
Pearson		**78.50**	<u>90.07</u>	**66.29**	<u>82.41</u>	<u>10.62</u>	<u>10.67</u>	<u>5.19</u>	<u>5.11</u>
baseline	Domain 3	83.49	**93.36**	<u>72.77</u>	**87.82**	<u>8.06</u>	<u>6.19</u>	<u>3.89</u>	**3.15**
JS		**84.82**	**93.36**	**74.66**	<u>87.53</u>	**7.56**	6.24	**3.62**	**3.15**
Jeffrey		<u>83.58</u>	93.12	72.71	87.40	8.78	<u>6.23</u>	4.00	<u>3.17</u>
Pearson		83.03	<u>93.14</u>	72.02	87.43	9.08	<u>6.23</u>	4.20	3.19
baseline	Domain 4	78.63	**93.56**	66.27	**88.16**	10.95	6.28	5.44	**3.06**
JS		**81.63**	92.18	**70.23**	85.87	**9.32**	8.01	**4.42**	3.87
Jeffrey		<u>79.24</u>	<u>93.31</u>	66.86	<u>87.74</u>	10.48	**6.15**	5.00	<u>3.16</u>
Pearson		80.03	93.12	<u>67.92</u>	87.43	<u>10.51</u>	6.31	<u>4.93</u>	3.28

Table 5. Performance comparison of different f-divergence strategies across four domains on Fundus dataset. Metrics marked with ↑ indicate that higher values imply better performance, while those with ↓ suggest the opposite. The best performance results are highlighted in bold, and the second-best are underlined.

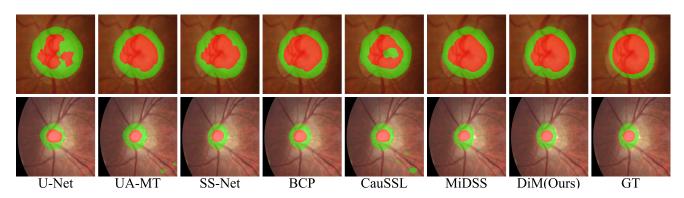| U-Net | UA-MT | SS-Net | BCP | CauSSL | MiDSS | DiM(Ours) | GT |

Figure 2. Visual comparison of segmentation results on Fundus dataset across different models. Red and green represent the Optical Cup and Disc, respectively. The first row shows segmentation results on test samples from the labeled domain (Domain 1), while the second row presents results on test samples from a different domain (Domain 4).

resentation obtained from our method, as shown in Figure. 1(a) and 1(b). DiM significantly improves the feature alignment between different domains compared to MiDSS, resulting in a tighter and more coherent data distribution.

Segmentation visualization. As shown in Figure 2, DiM achieves higher accuracy and better preservation of edge details in optic disc and cup segmentation tasks compared to other methods. Models such as U-Net, UA-MT, and SS-Net show noticeable boundary blurring, with segmentation results that lack precision, particularly at the structural boundaries of the optic disc and cup. These models tend to either miss segments or over-segment certain regions. In contrast, our model produces clearer boundaries with more precise edge detail retention, and the segmentation within the optic disc and cup regions is more cohesive, closely matching the ground truth (GT).

We also conducted visualization experiments on the prostate segmentation task in Figure 3. Results indicate that DiM continues to outperform other methods, such as BCP, CauSSL, and MiDSS, by achieving clearer boundary delineation and better edge detail preservation. The segmentation closely matches the GT, demonstrating improved accuracy and cohesion within the prostate region, even in challenging boundary areas.

Model Performance. The validation loss and Dice coefficient curves across the four domains on the Fundus dataset
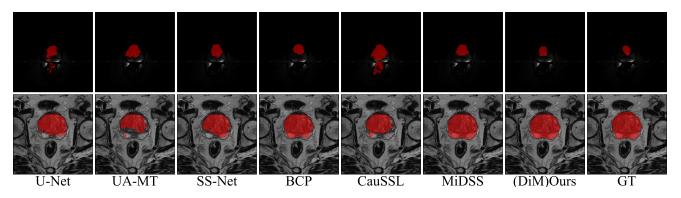
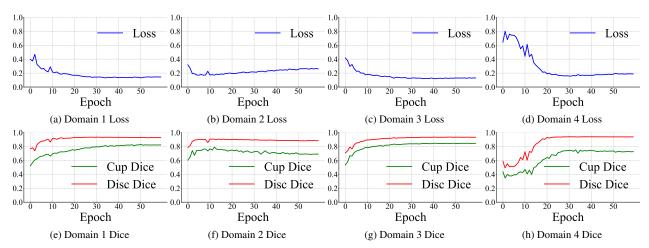Figure 3. Visual results on Prostate dataset.

Figure 4. Validation Loss and Dice across four domains on Fundus dataset.

demonstrate stable model performance, which are depicted in Figure 5. Loss decreases rapidly in the early epochs and converges across all domains, indicating effective training. Dice scores for both the optic cup and disc steadily increase and plateau at high values, with the optic disc achieving near-perfect accuracy.

5. Conclusion

This study addresses the challenge of data annotation in medical image segmentation by introducing a sharpness-aware optimization method based on f-divergence minimization (DiM) for semi-supervised learning. While existing semi-supervised methods (SSMIS), including sharpness-aware optimization (SAM), have shown success, they often overlook distribution differences between datasets. The proposed DiM method enhances model stability by adjusting the sensitivity of model parameters and improves adaptability to varying datasets. By reducing f-divergence, DiM achieves a better balance in performance between source and target datasets and mitigates overfitting. Experimental results demonstrate that DiM significantly improves performance, as evidenced by its ground-
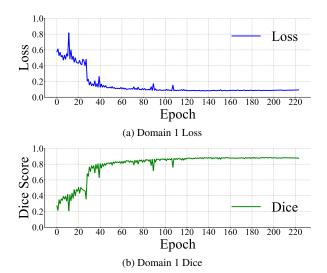

Figure 5. Validation Loss and Dice for Domain 1 on the Prostate dataset.

breaking progress in Dice scores on the prostate dataset, with similar success across three public datasets.

References

[1] Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *ICML*, pages 639–668. PMLR, 2022. 3

[2] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *CVPR*, pages 11514–11524, 2023. 5

[3] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen De Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *MICCAI*, pages 810–818. Springer, 2019. 1, 2

[4] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *TMI*, 39(7):2494–2505, 2020. 5

[5] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, pages 2613–2622, 2021. 5

[6] Dansong Cheng, Feng Tian, Lin Liu, Xiaofang Liu, and Ye Jin. Image segmentation based on multi-region multi-scale local binary fitting and kullback–leibler divergence. *Signal, Image and Video Processing*, 12:895–903, 2018. 2

[7] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1):1–25, 2019. 1

[8] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 1, 3

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189. PMLR, 2015. 2

[10] Joumana Ghosn and Yoshua Bengio. Bias learning, knowledge sharing. *TNN*, 14(4):748–765, 2003. 2

[11] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[13] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *TMI*, 37(12):2663–2674, 2018. 1

[14] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, pages 1013–1023, 2021. 1

[15] Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, Jonghye Woo, et al. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022. 2

[16] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *NeurIPS*, 31, 2018. 2

[17] Changjie Lu, Shen Zheng, and Gaurav Gupta. Unsupervised domain adaptation for cardiac segmentation: Towards structure mutual information maximization. In *CVPR*, pages 2588–2597, 2022. 5

[18] Liyun Lu, Mengxiao Yin, Liyao Fu, and Feng Yang. Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation. *Biomedical Signal Processing and Control*, 79:104203, 2023. 2

[19] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *AAAI*, pages 8801–8809, 2021. 2

[20] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *International conference on medical imaging with deep learning*, pages 820–833. PMLR, 2022. 2

[21] Qinghe Ma, Jian Zhang, Lei Qi, Qian Yu, Yinghuan Shi, and Yang Gao. Constructing and exploring intermediate domains in mixed domain semi-supervised medical image segmentation. In *CVPR*, pages 11642–11651, 2024. 1, 5

[22] Qinghe Ma, Jian Zhang, Lei Qi, Qian Yu, Yinghuan Shi, and Yang Gao. Constructing and exploring intermediate domains in mixed domain semi-supervised medical image segmentation. In *CVPR*, pages 11642–11651, 2024. 2

[23] Juzheng Miao, Cheng Chen, Furui Liu, Hao Wei, and Pheng-Ann Heng. Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In *ICCV*, pages 21426–21437, 2023. 1, 2, 5

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 1, 5

[25] Vivek A Rudrapatna, Atul J Butte, et al. Opportunities and challenges in using real-world data for health care. *The Journal of Clinical Investigation*, 130(2):565–574, 2020. 1

[26] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, pages 8050–8058, 2019. 2

[27] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 2

[28] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, pages 3752–3761, 2018. 2

[29] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *TMI*, 41(3):608–620, 2021. 5

[30] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018. 2

[31] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying

semi-supervised learning with consistency and confidence. *NeurIPS*, 33:596–608, 2020. 5

[32] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. 2

[33] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Medical Image Analysis*, 79:102447, 2022. 1

[34] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 2

[35] Qin Wang, Wen Li, and Luc Van Gool. Semi-supervised learning by augmented distribution alignment. In *ICCV*, pages 1466–1475, 2019. 1

[36] Xin Wang, Xiaoyu Liu, Peng Huang, Pu Huang, Shu Hu, and Hongtu Zhu. U-medsam: Uncertainty-aware medsam for medical image segmentation. *arXiv preprint arXiv:2408.08881*, 2024. 3

[37] Yiqing Wang, Zihan Li, Jieru Mei, Zihao Wei, Li Liu, Chen Wang, Shengtian Sang, Alan L Yuille, Cihang Xie, and Yuyin Zhou. Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In *MICCAI*, pages 486–496, 2023. 2

[38] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3:1–40, 2016. 2

[39] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022. 2

[40] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *MICCAI*, pages 34–43. Springer, 2022. 5

[41] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, pages 4085–4095, 2020. 5

[42] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *MICCAI*, pages 605–613. Springer, 2019. 5

[43] Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, and Mingkui Tan. Collaborative unsupervised domain adaptation for medical image diagnosis. *TIP*, 29:7834–7844, 2020. 1

[44] Ziyuan Zhao, Fangcheng Zhou, Kaixin Xu, Zeng Zeng, Cuntai Guan, and S Kevin Zhou. Le-uda: Label-efficient unsupervised domain adaptation for medical image segmentation. *TMI*, 42(3):633–646, 2022. 1

[45] Yixuan Zhou, Yi Qu, Xing Xu, and Hengtao Shen. Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In *ICCV*, pages 11345–11355, 2023. 3

[46] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *TMI*, 39(6):1856–1867, 2019. 1

[47] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018. 1

[48] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. *arXiv preprint arXiv:2203.08065*, 2022. 2

[49] Xiahai Zhuang et al. Challenges and methodologies of fully automatic whole heart segmentation: a review. *Journal of Healthcare Engineering*, 4:371–407, 2013. 2