DiffRect: Latent Diffusion Label Rectification for Semi-supervised Medical Image Segmentation

Xinyu Liu, Wuyang Li, and Yixuan Yuan✉

Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong SAR
yxyuan@ee.cuhk.edu.hk

Abstract. Semi-supervised medical image segmentation aims to leverage limited annotated data and rich unlabeled data to perform accurate segmentation. However, existing semi-supervised methods are highly dependent on the quality of self-generated pseudo labels, which are prone to incorrect supervision and confirmation bias. Meanwhile, they are insufficient in capturing the label distributions in latent space and suffer from limited generalization to unlabeled data. To address these issues, we propose a Latent Diffusion Label Rectification Model (DiffRect) for semi-supervised medical image segmentation. DiffRect first utilizes a Label Context Calibration Module (LCC) to calibrate the biased relationship between classes by learning the category-wise correlation in pseudo labels, then apply Latent Feature Rectification Module (LFR) on the latent space to formulate and align the pseudo label distributions of different levels via latent diffusion. It utilizes a denoising network to learn the coarse to fine and fine to precise consecutive distribution transportations. We evaluate DiffRect on three public datasets: ACDC, MS-CMRSEG 2019, and Decathlon Prostate. Experimental results demonstrate the effectiveness of DiffRect, e.g. it achieves 82.40% Dice score on ACDC with only 1% labeled scan available, outperforms the previous state-of-the-art by 4.60% in Dice, and even rivals fully supervised performance. Code is released at https://github.com/CUHK-AIM-Group/DiffRect.

Keywords: Semi-supervised · Medical Image Segmentation · Diffusion Models · Label Rectification.

1 Introduction

Medical image segmentation is crucial for clinical applications but often requires large amounts of pixel-wise or voxel-wise labeled data, which is tedious and time-consuming to obtain [17, 19, 18, 1, 28]. Such a heavy annotation cost has motivated the community to develop semi-supervised learning methods [10, 20, 14, 38]. Existing semi-supervised image segmentation methods can be generally categorized into self-training and consistency regularization. For self-training methods [1, 31, 4, 7, 34, 15, 40, 23], they generate pseudo labels for unlabeled images, then use the pseudo-labeled images in conjunction with labeled images to update the segmentation model iteratively. This paradigm could effectively incorporate unlabeled data by minimizing their entropy. For consistency regularization methods [5, 27, 21, 30, 35, 9, 22, 37, 33], they are designed based on the assumption that perturbations should not change the predictions of the model, and have achieved more promising performance recently. Perturbations are applied on the input or the network level, and the models are enforced to achieve an invariance of predictions.

Despite the progress, the semi-supervised medical image segmentation remains challenging due to the following factors. (1) **Reliance Risk**: Existing methods typically rely on self-generated pseudo labels to optimize the model [30, 37, 11, 35], which is ill-posed since errors in pseudo labels are preserved during iterative optimization. The overfitting to incorrect supervision could lead to severe confirmation bias [16] and considerable performance degradation. Besides, they do not fully utilize the category-wise correlation in the pseudo labels, and the label quality is sensitive to the perturbation design and network structure. (2) **Distribution Misalignment**: Most methods only apply consistency regularization and auxiliary supervision at the output mask level to encourage the model to produce consistent mask predictions between different perturbations [27, 5]. However, these approaches are insufficient in capturing the semantics in the latent space and tend to overlook the underlying label distributions, resulting in limited generalization to unlabeled data.

To address the reliance risk issue, we first propose a **L**abel **C**ontext **C**alibration Module (LCC). Different from methods that directly use the self-generated pseudo labels, LCC calibrates the biased semantic context, *i.e.*, the relationships between different semantic categories, and reduce the errors in the pseudo labels. It starts with a semantic coloring scheme that encodes the one-hot pseudo labels and ground truth masks into the visual space, and subsequently feeds them into a semantic context embedding block to adjust the features of the pseudo labels in the latent space. Notably, LCC introduces explicit calibration guidance by encoding the dice score between the pseudo labels and the ground truth, thereby providing more reliable calibration directions for model optimization.

To tackle the distribution misalignment problem, some previous works have proposed to model data distributions with VAE [41] or GAN [42]. However, their adversarial training scheme could suffer from mode collapse and conflict between generation and segmentation tasks, resulting in suboptimal performance. Different from them, the denoising diffusion probabilistic model (DDPM) is a new class of generative models trained using variational inference [8, 24, 12, 13], which alleviates the above problem by formulating the complex data distribution with probabilistic models. Therefore, we design a **L**atent **F**eature **R**ectification Module (LFR), which models the consecutive refinement between different latent distributions with a generative latent DDPM [25]. LFR leverages the power of DDPM to learn the latent structure of the semantic labels. Specifically, it first applies Gaussian noise on fine-grained label features with a diffusion schedule, then uses the coarse-grained label features as conditions to recover the clean feature. With the denoising process, the consecutive transportations of coarse to fine and fine to precise distributions of the pseudo labels are formulated and aligned, and the pseudo labels are progressively rectified for better supervision. Based on LCC and LFR, we construct a semi-supervised medical image segmentation framework named Latent **Diff**usion Label **Rect**ification Model (*DiffRect*). Extensive experimental results show that our method outperforms prior methods by significant margins.

2 Methodology

2.1 Preliminary: Conditional DDPM

DDPM is a class of latent variable generative model that learns a data distribution by denoising noisy images [8]. The forward process diffuses the data samples with pre-defined noise schedules. Concretely, given a clean data z^0, sampling of z^t is expressed in a closed form:

$$q(z^t \| z^0) = \mathcal{N}(z^t; \sqrt{\overline{\alpha}_t} z^0, (1-\overline{\alpha}_t)\mathbf{I}), \tag{1}$$

where $\overline{\alpha}_t$ is the noise schedule variable [24, 8]. During the reverse process, we are given an optional condition ρ [6], and each step is expressed as a Gaussian transition with learned mean $\boldsymbol{\mu}_\epsilon$ and variance σ_ϵ from the denoising model ϵ:

$$p\left(z^{t-1} \mid z^t, \rho\right) := \mathcal{N}\left(z^{t-1}; \boldsymbol{\mu}_\epsilon\left(z^t, t, \rho\right), \sigma_\epsilon\left(z^t, t, \rho\right)\mathbf{I}\right). \tag{2}$$

By decomposing the above equation, we have:

$$z^{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}(z^t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon(z^t, t, \rho)) + \sigma_\epsilon \eta, \tag{3}$$

where $\eta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a sampled noise that ensures each step is stochastic. In this work, we extend the conditional DDPM to the latent space of pseudo labels, and model the distribution transportations for label rectification.

2.2 Label Context Calibration Module (LCC)

Existing semi-supervised training schemes that rely extensively on self-generated pseudo labels are often ill-posed, where errors in low-quality pseudo labels accumulate and degrade performance. To address this issue, we introduce LCC that effectively captures and calibrates the semantic context within the visual space, thereby mitigating the impact of noisy labels. As in Fig. 1(a), given the one-hot pseudo labels $y_s, y_w \in \mathbb{R}^{H\times W\times C}$ with height H and width W from the segmentation network, we encode them to semantic pseudo labels m_s and m_w with dimensions of $\mathbb{R}^{H\times W\times 3}$, using a proposed *semantic coloring scheme (SCS)*.

Concretely, for a dataset that contains C different classes, we build a color set M_C that is composed of C RGB colors, and each color is represented by a tuple of three values within the range $[0, 255]$. We maximize the color difference between each encoded category to avoid semantic confusion. Therefore, it can be represented by a functional mapping $f: C \to M_C$, which is defined as:

$$m_{(h,w)} = f(y_{(h,w)}), \quad \forall h \in [1, 2, ..., H], w \in [1, 2, ..., W], \tag{4}$$

where m is the semantic pseudo label in the visual space, and $m_{(h,w)}$ represents the mapped RGB color of the pixel at location (h, w) in m. The $y_{(h,w)}$

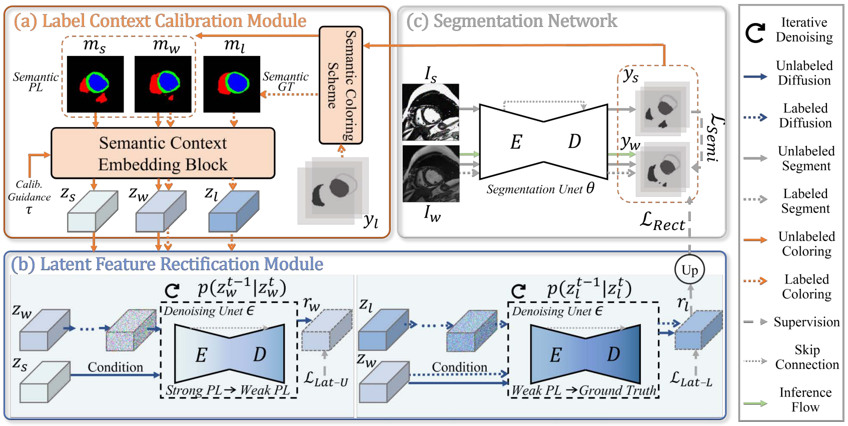

Fig. 1. Overall framework of DiffRect. (a) Label Context Calibration Module (LCC). (b) Latent Feature Rectification Module (LFR). (c) Segmentation Network.

represents the class of the corresponding pixel in one-hot mask y. The semantic coloring scheme can effectively incorporate color information into the segmentation task, which enables the model to exploit additional cues with the rich semantics from colors, and improves the discrimination ability [32, 3] as well as the interpretability of the model.

To perform context calibration with the semantic labels, we design a semantic context embedding block $\mathbf{B}_{sem}$, which embeds the pseudo labels to the latent features z_s, z_w, z_l with the dimensions of $\mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 256}$. Specially, additional *calibration guidance (CG)* τ^u for unlabeled data and τ^l for labeled data are also encoded into the block using the sinusoidal embeddings [8, 29],

$$\begin{aligned} \{z_s, z_w\} &= \mathbf{B}_{\text{sem}}(m_s, m_w \| \tau^u) \quad \text{for unlabeled data,} \\ \{z_w, z_l\} &= \mathbf{B}_{\text{sem}}(m_w, m_l \| \tau^l) \quad \text{for labeled data,} \end{aligned} \tag{5}$$

where the τ^u and τ^l values for unlabeled and labeled data are computed using the dice coefficient between the one-hot segmentation masks of different qualities, which is denoted as follows:

$$\tau^u = \text{Dice}(y_s, y_w), \quad \tau^l = \text{Dice}(y_w, y_l). \tag{6}$$

By using the dice coefficient as the calibration guidance factor, the model can simultaneously measure the quality of pseudo labels and integrate this information into the learning process. It enables the model to better capture the semantic context and refine the pseudo labels for both unlabeled and labeled data.

2.3 Latent Feature Rectification Module (LFR)

To address the distribution misalignment issue between the pseudo labels with different levels of quality, we propose a Latent Feature Rectification Module (LFR), which is illustrated in Fig. 1(b).

Concretely, LFR applies a latent diffusion process to model the transportation of label quality distributions. For each unlabeled data I_u, the strongly and weakly semantic context embedding z_s and z_w are first obtained with LCC. We then construct a diffusion process from z_w to the diffused noisy feature z_w^T with T timestamps as follows:

$$\begin{aligned} z_w^T &= \sqrt{\alpha_T} z_w^{T-1} + \sqrt{1-\alpha_T}\eta^{T-1} \\ &= \cdots = \sqrt{\overline{\alpha}_T} z_w + \sqrt{1-\overline{\alpha}_T}\eta, \end{aligned} \tag{7}$$

where α_T and $\overline{\alpha}_T$ are the schedule variables in the diffusion forward process, (*e.g.*, cosine [24]), and $\overline{\alpha}_T = \prod_{i=1}^{T} \alpha_i$. The η^t is the corresponding noise sampled from Gaussian distribution at the t-th step. Then, we train a denoising U-Net ϵ to learn to reverse this process. Since the individual reverse diffusion process is unconditioned, we add z_s as the conditional input and also feed it into the denoising model. Therefore, the model is encouraged to learn the distribution transportation from *coarse-grained masks* $p(z_s)$ *(strong pseudo labels)* to the latent distributions of *fine-grained masks* $p(z_w)$ *(weak pseudo labels)*, where we denote it as a *strong-to-weak transportation (S2W)*. The reverse diffusion is formulated as the following Markov chain:

$$\begin{aligned} & p_\epsilon\left(z_w^{0:T}\right) := p\left(z_w^T\right) \prod_{t=1}^{T} p_\epsilon\left(z_w^{t-1} \mid z_w^t, z_s\right), \quad z_w^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ & p_\epsilon\left(z_w^{t-1} \mid z_w^t, z_s\right) := \mathcal{N}\left(z_w^{t-1}; \boldsymbol{\mu}_\epsilon\left(z_w^t, t, z_s\right), \sigma_\epsilon\left(z_w^t, t, z_s\right)\mathbf{I}\right), \end{aligned} \tag{8}$$

where $\boldsymbol{\mu}$ and σ are the predicted data mean and variance from the denoising U-Net model. For the training with unlabeled input, the *latent loss* for optimization can be expressed as follows:

$$\mathcal{L}_{\text{Lat-U}} = E_{z_w,t}\left[\|z_w - r_w\|_2\right], \tag{9}$$

where $r_w = \epsilon\left(z_w^T, z_s, t\right)$, which is the reconstructed version of the weakly semantic context embedding z_w. The objective minimizes the ℓ_2 distance between the clean and denoised feature and encourages the model to learn the distribution transportation from a coarse pseudo label to a fine pseudo label.

Similarly, we can obtain the weak semantic context embedding of labeled data z_w and the ground truth z_l. We then learn the reverse process that recovers z_l based on the T-timestamp diffused noisy feature z_l^T, with the z_w as condition:

$$\begin{aligned} & p_\epsilon\left(z_l^{0:T}\right) := p\left(z_l^T\right) \prod_{t=1}^{T} p_\epsilon\left(z_l^{t-1} \mid z_l^t, z_w\right), \quad z_l^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ & p_\epsilon\left(z_l^{t-1} \mid z_l^t, z_w\right) := \mathcal{N}\left(z_l^{t-1}; \boldsymbol{\mu}_\epsilon\left(z_l^t, t, z_w\right), \sigma_\epsilon\left(z_l^t, t, z_w\right)\mathbf{I}\right), \end{aligned} \tag{10}$$

and the training objective for the reconstructed feature $r_l = \epsilon\left(z_l^T, z_w, t\right)$ is:

$$\mathcal{L}_{\text{Lat-L}} = E_{z_l,t}\left[\|z_l - r_l\|_2\right]. \tag{11}$$

With the above latent diffusion process, the continual distribution transportations from *fine-grained mask distributions* $p(z_w)$ *(weak pseudo labels)* to *precise* *mask distributions* $p(z_l)$ *(ground truth)* are also formulated in the latent space, which is denoted as the *weak-to-groud truth transportation (W2G)*. The denoising U-Net is hence capable to achieve latent feature rectification.

Afterwards, the weak pseudo labels of unlabeled data are fed into the denoising U-Net for obtaining the rectified features with progressive denoising. Specifically, we randomly sample a Gaussian noise $r_l^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as the input of the denoising U-Net, which simulates the T-timestamp noisy feature of the rectified pseudo label y_r. The rectified feature r_l is generated via a progressive reverse diffusion process, with the weak pseudo label features z_w as condition. Mathematically, a single denoising from step t to $t-1$ is formulated as:

$$r_l^{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}(r_l^t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon(r_l^t, t, z_w)) + \sigma_\epsilon \eta, \tag{12}$$

where $\eta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ which ensures each step is stochastic as in DDPM [8]. The rectified label is obtained with an upsampling of the feature r_l to the input resolution $y_r = \mathit{Upsample}(r_l)$, which is utilized as a better and more precise supervision signal for the segmentation model.

2.4 Loss Function

The training of the DiffRect frameworks includes two parts: (1) the optimization of segmentation U-Net θ (with Seg Loss) and (2) the joint optimization of the rectification components $\mathbf{B}_{sem}$ and ϵ (with Diff Loss). The overall loss is:

$$\mathcal{L}_{\text{DiffRect}} = \underbrace{\mathcal{L}_{\text{Semi}}^{\text{Seg}} + \mathcal{L}_{\text{Rect}}}_{\text{Seg Loss}} + \underbrace{\mathcal{L}_{\text{Semi}}^{\text{Lat}} + \lambda_1 \mathcal{L}_{\text{Lat-U}} + \lambda_2 \mathcal{L}_{\text{Lat-L}}}_{\text{Diff Loss}}, \tag{13}$$

where $\mathcal{L}_{\text{Semi}}^{\text{Seg}}$ and $\mathcal{L}_{\text{Semi}}^{\text{Lat}}$ are the semi-supervised losses for segmentation as in [27]. The λ_1 and λ_2 are trade-off factors to balance the contribution of each term. $\mathcal{L}_{\text{Rect}}$ is the rectified supervision loss between y_w and the rectified pseudo label y_r, where the summation of cross-entropy and Dice score are used:

$$\mathcal{L}_{\text{Rect}} = \text{CE}(y_w, y_r) + \text{Dice}(y_w, y_r). \tag{14}$$

During inference, the input is directly fed into segmentation network in Fig. 1(c) to produce the segmentation result, thus no extra inference cost is required.

3 Experiments

3.1 Experimental Setup

We examine all methods with identical settings for fair comparison, and trained on a NVIDIA 4090 GPU for 30k iterations. For the $\mathbf{B}_{sem}$ which downsamples the input to $\frac{H}{16} \times \frac{W}{16}$, we use two 3×3 convolution layers followed by BN and LeakyReLU before the 2× downsample in each stage, and repeat for four stages. The Denoising U-Net ϵ down and upsamples the input by 4×, which also uses two 3×3 convolution layers per stage. The multi-scale image feature is embedded into the model via concatenation as in [36]. For the weak perturbation, we apply

Table 1. Segmentation results on the ACDC validation and test sets.

Method	Labeled Ratio	ACDC Validation Set				ACDC Test Set			
		Dice↑	Jac↑	HD95↓	ASD↓	Dice↑	Jac↑	HD95↓	ASD↓
UAMT [39]	1%	42.28	32.21	40.74	18.58	43.86	33.36	38.60	18.33
FixMatch [27]		69.67	58.34	37.92	14.41	60.80	49.14	36.81	14.75
CPS [5]		56.70	44.31	24.97	10.48	52.28	41.68	20.38	7.35
ICT [30]		43.03	30.58	34.92	15.23	42.91	32.81	25.42	10.80
MCNetV2 [35]		57.49	43.29	31.31	10.97	49.92	39.16	24.64	8.47
INCL [43]		77.80	66.13	11.69	3.22	67.01	56.22	13.43	3.35
DiffRect (Ours)		**82.40**	**71.96**	**10.04**	**2.90**	**71.85**	**61.53**	**5.79**	**2.12**
UAMT [39]	5%	72.71	60.89	21.48	7.15	69.93	58.45	17.01	5.25
FixMatch [27]		83.12	73.59	9.86	2.61	74.68	64.12	11.18	2.93
CPS [5]		75.24	64.67	10.93	2.98	74.67	63.51	9.37	2.55
ICT [30]		74.20	62.90	17.01	4.32	73.10	60.69	11.92	3.70
MCNetV2 [35]		78.96	68.15	12.13	3.91	75.86	65.20	9.85	2.88
INCL [43]		85.43	75.76	6.37	1.37	80.64	70.78	**5.29**	**1.42**
DiffRect (Ours)		**86.95**	**78.08**	**4.07**	**1.23**	**82.46**	**71.76**	7.18	1.94
UAMT [39]	10%	85.14	75.90	6.25	1.80	86.23	76.72	9.40	2.56
FixMatch [27]		88.31	79.97	7.35	1.79	87.96	79.37	5.43	1.59
CPS [5]		84.63	75.20	7.57	2.27	85.61	75.76	9.29	3.00
ICT [30]		85.15	76.05	4.27	1.46	86.77	77.43	8.01	2.16
MCNetV2 [35]		85.97	77.21	7.55	2.11	88.75	80.28	6.16	1.64
INCL [43]		88.28	80.09	1.67	0.49	88.68	80.27	4.34	1.13
DiffRect (Ours)		**90.18**	**82.72**	**1.38**	**0.48**	**89.27**	**81.13**	**3.85**	**1.00**
Supervised [26]	100%	91.48	84.87	1.12	0.34	91.65	84.95	1.14	0.50

random flipping and rotation. For the strong perturbation, we apply random Gaussian blur and additional random image adjustments, including contrast, sharpness, and brightness enhancement. For ACDC, we test the 1%, 5%, and 10% labeling regimes following [20]. For MS-CMRSEG 2019, 20% labeling regime is tested, while 10% labeled data is used in Decathlon Prostate.

3.2 Comparison with State-of-the-art Methods

We validate the effectiveness of the proposed approach on the ACDC dataset [2] in Tab. 1. Our method shows superior results under all labeling regimes. Compared with MCNetV2 [35], our method possesses superior capability with increments of 24.91%, 7.99%, 4.21% in Dice, 28.67%, 9.93%, 5.51% in Jaccard on the validation set with 1%, 5%, and 10% scans available. DiffRect displays better segmentation performance even when the labeled samples are extremely scarce (*e.g.* 82.40% Dice with 1% scans available), suggesting it can model the transportation of the pseudo label distributions precisely and produce refined masks. Results in MS-CMRSEG 2019 are shown in Tab. 2. DiffRect shows consistent performance gain on all metrics, with 86.78% in Dice, 77.13% in Jaccard, 6.39mm in HD95, and 1.85mm in ASD, outperforming the state-of-the-art method INCL [43] by 2.45% Dice, 3.21% Jaccard, 3.56mm HD95, and 0.76mm in ASD, respectively. On Decathlon Prostate in Tab. 3, DiffRect remains showing compelling results, demonstrating its capability in various modalities.

Table 2. Segmentation results on MS-CMRSEG 2019 with 20% data labeled.

Method	Dice $\uparrow$	Jac$\uparrow$	HD95$\downarrow$	ASD$\downarrow$
UAMT [39]	84.27	73.69	12.15	4.18
FixMatch [27]	84.31	73.57	17.79	4.81
CPS [5]	83.66	73.03	15.01	4.30
ICT [30]	83.66	73.06	17.24	4.85
MCNetV2 [35]	83.93	73.45	13.10	3.39
INCL [43]	84.33	73.92	9.95	2.61
DiffRect	**86.78**	**77.13**	**6.39**	**1.85**
Supervised [26]	88.19	79.28	4.21	1.32

Table 3. Segmentation results on Decathlon Prostate with 10% data labeled.

Method	Dice$\uparrow$	Jac$\uparrow$	HD95$\downarrow$	ASD$\downarrow$
UAMT [39]	40.91	29.13	28.32	10.45
FixMatch [27]	54.70	41.07	16.82	5.24
CPS [5]	43.51	31.18	26.93	8.31
ICT [30]	39.91	28.95	24.73	7.59
MCNetV2 [35]	40.58	28.77	21.29	7.11
INCL [43]	55.67	41.91	31.09	15.78
DiffRect	**62.23**	**48.64**	**10.36**	**3.41**
Supervised [26]	73.81	61.25	7.28	1.94

Table 4. Ablation study of the proposed modules.

Method	w/o	Dice$\uparrow$	Jac$\uparrow$	HD95$\downarrow$	ASD$\downarrow$
Baseline	-	69.67	58.34	37.92	14.41
+LCC	SCS	73.83	61.83	29.49	11.71
	CG	76.12	64.69	26.24	8.31
	-	78.28	66.97	20.46	5.60
+LCC & LFR	S2W	79.97	69.31	14.07	4.91
	W2G	78.57	66.38	21.07	5.91
	-	**82.40**	**71.96**	**10.04**	**2.90**

Table 5. Ablation study of different calibration guidance choices in LCC.

Choice	Dice$\uparrow$	Jac$\uparrow$	HD95$\downarrow$	ASD$\downarrow$
Dice	**82.40**	**71.96**	**10.04**	2.90
Jaccard	82.37	71.82	11.33	2.87
Fixed	80.34	69.67	14.97	4.47
Random	80.60	69.99	13.15	3.75
Both	81.67	71.45	10.28	**2.47**

3.3 Further Analysis

Ablation study of the proposed modules. We evaluate the effect of individual modules in DiffRect in Tab. 4. Adopting LCC achieves 78.28% Dice and 66.97% Jaccard, with 8.61% and 8.63% gains compared with the Fixmatch baseline [27]. Removing the semantic coloring scheme (SGS) shows a large performance drop (73.83% Dice and 61.83% Jaccard), showing the importance of exploiting the semantics in the visual domain. No calibration guidance (CG) causes 2.16% Dice drop due to the impact of noisy calibration directions. Adding LFR improves Dice by 4.12% and 10.42mm in HD95. Removing the strong to weak transportation (S2W) shows a 2.43% Dice drop while removing the weak to ground truth (W2G) causes a severe Dice drop to 78.57%. The results demonstrate the necessity of each sub-component.

Different Calibration Guidance Choices. To analyze the effectiveness and the optimal choice of calibration guidance, experiments were conducted to compare the performance of models trained with different calibration guidance in Tab. 5, including Dice score, Jaccard score, Fixed (using a fixed value 0.5), Random (using a random sampled value within 0$\sim$1), and Both (using the summation of Dice and Jaccard). It is shown that Dice, Jaccard, and Both have similar performance, and outperform the fixed and random strategies, which validates the reliable directions provided for optimization.

4 Conclusion

In this paper, we identify the reliance risk and distribution misalignment issues in semi-supervised medical image segmentation, and propose DiffRect, a diffusion-based framework for this task. It comprises two modules: the LCC aims to calibrate the biased relationship between classes in pseudo labels by learning category-wise correlation, and the LFR models the consecutive transportations between coarse to fine and fine to precise distributions of the pseudo labels accurately with latent diffusion. Extensive experiments on three datasets demonstrate that DiffRect outperforms existing methods by remarkable margins.

Acknowledgments. This work was supported by Hong Kong Research Grants Council (RGC) General Research Fund 14204321.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac mr image segmentation. In: MICCAI. pp. 253–260. Springer (2017)
2. 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.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
3. Chen, J., Lu, J., Zhu, X., Zhang, L.: Generative semantic segmentation. In: CVPR. pp. 7111–7120 (2023)
4. Chen, S., Bortsova, G., García-Uceda Juárez, A., Van Tulder, G., De Bruijne, M.: Multi-task attention-based semi-supervised learning for medical image segmentation. In: MICCAI. pp. 457–465. Springer (2019)
5. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR. pp. 2613–2622 (2021)
6. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. In: ICCV. pp. 14347–14356. IEEE (2021)
7. Feng, Z., Zhou, Q., Cheng, G., Tan, X., Shi, J., Ma, L.: Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum. arXiv preprint arXiv:2004.08514 **1**(2), 5 (2020)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
9. Hu, H., Wei, F., Hu, H., Ye, Q., Cui, J., Wang, L.: Semi-supervised semantic segmentation via adaptive equalization learning. NeurIPS **34**, 22106–22118 (2021)
10. Jiao, R., Zhang, Y., Ding, L., Cai, R., Zhang, J.: Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. arXiv preprint arXiv:2207.14191 (2022)
11. Li, C., Lin, M., Ding, Z., Lin, N., Zhuang, Y., Huang, Y., Ding, X., Cao, L.: Knowledge condensation distillation. In: ECCV. pp. 19–35 (2022)
12. Li, C., Liu, H., Liu, Y., Feng, B.Y., Li, W., Liu, X., Chen, Z., Shao, J., Yuan, Y.: Endora: Video generation models as endoscopy simulators. arXiv preprint arXiv:2403.11050 (2024)

13. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Yuan, Y.: U-kan makes strong backbone for medical image segmentation and generation. arXiv:2406.02918 (2024)
14. Li, C., Ma, W., Sun, L., Ding, X., Huang, Y., Wang, G., Yu, Y.: Hierarchical deep network with uncertainty-aware semi-supervised learning for vessel segmentation. NCA pp. 1–14 (2022)
15. Li, C., Zhang, Y., Liang, Z., Ma, W., Huang, Y., Ding, X.: Consistent posterior distributions under vessel-mixing: a regularization for cross-domain retinal artery/vein classification. In: ICIP. pp. 61–65. IEEE (2021)
16. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. In: ICLR (2019)
17. Liu, X., Guo, X., Liu, Y., Yuan, Y.: Consolidated domain adaptive detection and localization framework for cross-device colonoscopic images. Medical image analysis **71**, 102052 (2021)
18. Liu, X., Li, W., Yuan, Y.: Decoupled unbiased teacher for source-free domain adaptive medical object detection. IEEE Trans. Neural Netw. Learn. Syst. (2023)
19. Liu, X., Yuan, Y.: A source-free domain adaptive polyp detection framework with style diversification flow. IEEE Transactions on Medical Imaging **41**(7), 1897–1908 (2022)
20. Luo, X.: SSL4MIS. `https://github.com/HiLab-git/SSL4MIS` (2020)
21. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: MIDL. pp. 820–833. PMLR (2022)
22. Luo, X., Wang, G., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Metaxas, D.N., Zhang, S.: Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. Med. Image Anal. **80**, 102517 (2022)
23. Mendel, R., Rauber, D., de Souza Jr, L.A., Papa, J.P., Palm, C.: Error-correcting mean-teacher: Corrections instead of consistency-targets applied to semi-supervised medical image segmentation. CIBM **154**, 106585 (2023)
24. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML. pp. 8162–8171. PMLR (2021)
25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
27. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. NeurIPS **33** (2020)
28. Sun, L., Li, C., Ding, X., Huang, Y., Chen, Z., Wang, G., Yu, Y., Paisley, J.: Few-shot medical image segmentation using a global correlation network with discriminative embedding. CBM **140**, 105067 (2022)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
30. Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. Neural Netw. **145**, 90–106 (2022)
31. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR. pp. 2517–2526 (2019)
32. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: CVPR. pp. 6830–6839 (2023)

33. Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X.: Mcf: Mutual correction framework for semi-supervised medical image segmentation. In: CVPR. pp. 15651–15660 (2023)
34. Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., Le, X.: Semi-supervised semantic segmentation using unreliable pseudo-labels. In: CVPR. pp. 4248–4257 (2022)
35. Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J.: Mutual consistency learning for semi-supervised medical image segmentation. MIA **81**, 102530 (2022)
36. Xing, Z., Wan, L., Fu, H., Yang, G., Zhu, L.: Diff-unet: A diffusion embedded network for volumetric segmentation. arXiv preprint arXiv:2303.10326 (2023)
37. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: CVPR. pp. 7236–7246 (2023)
38. Yang, Q., Liu, X., Chen, Z., Ibragimov, B., Yuan, Y.: Semi-supervised medical image classification with temporal knowledge-aware regularization. In: MICCAI. pp. 119–129. Springer (2022)
39. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: MICCAI. pp. 605–613. Springer (2019)
40. Zhang, R., Liu, S., Yu, Y., Li, G.: Self-supervised correction learning for semi-supervised biomedical image segmentation. In: MICCAI. pp. 134–144. Springer (2021)
41. Zhang, X., Yao, L., Yuan, F.: Adversarial variational embedding for robust semi-supervised learning. KDD (2019)
42. Zhang, Y., Li, C., Lin, X., Sun, L., Zhuang, Y., Huang, Y., Ding, X., Liu, X., Yu, Y.: Generator versus segmentor: Pseudo-healthy synthesis. In: MICCAI. pp. 150–160 (2021)
43. Zhu, Y., Yang, J., Liu, S., Zhang, R.: Inherent consistent learning for accurate semi-supervised medical image segmentation. In: MIDL (2023)