A Dual-space Patch-driven Complementary Learning Framework for Semi-supervised Multi-organ Segmentation

Baixi Liang, Shuohong Xia, Sihao Li, and Yunyun Yang·*

Harbin Institute of Technology, Shenzhen, China
yangyunyun@hit.edu.cn

Abstract. Semi-supervised medical image segmentation (SSMIS) aims to achieve accurate segmentation with limited labeled data. However, existing methods still struggle in multi-organ scenarios. Limited annotations exacerbate class imbalance and feature entanglement among co-occurring organs, leading to false activations and performance degradation. To overcome these challenges, we propose a dual-space patch-driven complementary learning framework for semi-supervised multi-organ segmentation (DPCL). Our framework addresses these problems from both image and feature space. In the image space, we combine uncertainty and class frequency to form a probabilistic sampling strategy for image decomposition, which generates informative patches to enhance local feature perception and promote balanced learning across organs. In the feature space, we design two complementary objectives to achieve robust feature disentanglement. The patch-based contrastive loss focuses on decoupling features of co-occurring neighboring organs within local patches, mitigating semantic confusion among adjacent structures. Meanwhile, the global-based contrastive loss minimizes the gap between global and local semantics, improving model sensitivity to fine-grained details and enhancing consistency across different regions. These complementary losses collaboratively enhance inter-organ feature discrimination while mitigating false activations. Extensive evaluations on four public multi-organ medical segmentation datasets (Synapse, FeTA2021, SegTHOR, and ACDC) demonstrate that DPCL outperforms state-of-the-art methods, achieving substantial improvements in segmentation performance.

Keywords: Semi-supervised Learning · Medical Image Segmentation · Class Imbalanced · Feature Confusion

1 Introduction

Computed tomography (CT) and magnetic resonance imaging (MRI) have become indispensable pillars of modern clinical practice, providing non-invasive

* Corresponding author.

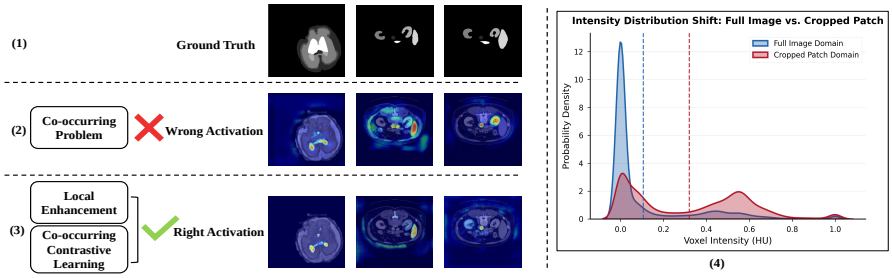


Fig. 1: Visualization of CAMs and intensity distributions. (1) Ground truth segmentation. (2) Co-occurrence issue: in semi-supervised segmentation, multi-organ co-occurrence leads to falsely activated regions, shown for ventricle, liver, and aorta. (3) CAMs after applying our local enhancement and co-occurring contrastive learning, which clearly reduce false activations and improve localization. (4) Comparison of intensity distributions between the full image and the cropped region.

insights that are critical for disease diagnosis, treatment planning, and intra-operative navigation [32]. In the era of precision medicine, the accurate segmentation of anatomical structures and organs is a prerequisite for downstream tasks such as radiation therapy and volumetric analysis. However, manual delineation by clinical experts is not only labor-intensive and time-consuming but also prone to inter-observer variability. While deep learning has revolutionized medical image segmentation, most state-of-the-art models rely heavily on fully supervised learning, which necessitates massive, high-quality voxel-wise annotations [7, 12, 19]. In the medical domain, acquiring such datasets is exceptionally challenging due to the scarcity of expert resources and the inherent complexity of 3D anatomical structures. Consequently, Semi-Supervised Medical Image Segmentation (SSMIS) has emerged as a promising paradigm, aiming to achieve robust performance by leveraging a small set of labeled data alongside a substantial volume of unlabeled images. Existing SSMIS frameworks primarily follow two technical routes: consistency regularization [2, 10, 36, 38], which enforces prediction stability under various perturbations, and pseudo-labeling [5, 30, 39], which iteratively expands the supervision signal using high-confidence model predictions. Despite these advancements, multi-organ segmentation in a semi-supervised setting remains a formidable challenge. The complexity arises from two intertwined issues. First, SSMIS inherently exacerbates class and scale imbalance. Small or underrepresented organs (e.g., the pancreas or gallbladder) are often overshadowed by dominant structures (e.g., the liver), leading to severe bias and high model uncertainty for minority classes. While recent studies have attempted to mitigate this via loss-weight rebalancing [1, 41] or specialized architectural designs [33, 40], they often overlook a more subtle yet destructive factor: Inter-class Co-occurrence. In clinical scans, certain organs consistently appear in close spatial proximity, such as the heart and aorta, or the liver and stomach. We observe that in data-scarce scenarios, this co-occurrence frequently leads to feature confusion. In scenarios with organ imbalance, the co-occurrence

of minority organs with dominant ones leads to feature confusion, resulting in false activations [8, 35, 37]. This mutual interference degrades the segmentation performance of both the minority and dominant organs. This issue is especially prevalent in semi-supervised segmentation, where the scarcity of labeled data further exacerbates the problem, as illustrated in Fig. 1(2). While prior work has made strides, it has largely focused on addressing class imbalance and the performance bottlenecks of SSMIS without directly tackling this co-occurrence challenge.

To address these limitations, we aim to tackle these challenges from both the image space and feature space. Consequently, we propose **DPCL**, a novel SSMIS framework designed to simultaneously refine local perception and decouple co-occurring organ features, as shown in Fig. 2. Our core motivation is that by forcing the model to focus on high-uncertainty local regions and explicitly enlarging the distance between confusing organ representations in the latent space, we can achieve more discriminative segmentation. By explicitly enlarging the feature discrepancy between different categories, our approach seeks to enhance the segmentation accuracy for all organ classes. The framework consists of two key modules: **Local Enhancement** and **Co-occurrence Decoupling**.

The **Local Enhancement** module leverages a novel decomposition strategy based on uncertainty and class frequencies. By sampling local patches from the image, this approach applies consistency regularization, which not only strengthens the model's segmentation performance on low-resolution or ambiguous regions but also improves its perception of these regions. This method is particularly effective in enhancing segmentation for boundary regions and underrepresented classes.

The **Co-occurrence Decoupling** module focuses on addressing the disorder of organ features that the local enhancement module fails to resolve, thereby further improving the segmentation performance for each organ. Consequently, based on the patches obtained from the local enhancement module, we first extract category-level knowledge from them. To further enhance the model's ability to separate features of co-occurring organs, we design a contrastive learning mechanism based on both patch-level and global-level feature representations. This approach effectively decouples the features of co-occurring organs, reducing semantic confusion and significantly improving segmentation quality. Our contributions are summarized as follow:

- We introduce DPCL, a new semi-supervised medical image segmentation framework that enhances the model's ability to perceive local regions and decouple features of co-occurring organs.
- We propose a novel Local Enhancement module driven by a dual-prior decomposition strategy. Besides, we design a joint local-and-global fusion mechanism to enhance the model's discriminative perception of localized structures, significantly improving segmentation robustness for underrepresented organs and complex boundary cases.

- We design a patch-based and global-based contrastive learning mechanism within the co-occurrence decoupling module, effectively addressing feature confusion caused by co-occurring organs.
- We validate the effectiveness of DPCL on four challenging medical segmentation datasets, including Synapse, Feta2021, SegTHOR, and ACDC, demonstrating substantial improvements in segmentation accuracy compared to existing state-of-the-art semi-supervised methods.

2 Related Work

2.1 Semi-supervised Medical Image Segmentation

In recent years, several state-of-the-art methods have been proposed for semi-supervised medical image segmentation [2, 10, 14, 15, 23, 33, 39, 42]. For instance, UA-MT [39] introduces the Mean-Teacher framework, which leverages uncertainty to enable the student model to learn meaningful information from the teacher model. Luo [23] propose a co-training framework that utilizes CNNs and Swin-Transformer [22] to generate high-quality pseudo-labels. BCP [2] and ABD [10] align labeled and unlabeled data distributions using bidirectional copy-paste and displacement strategies, respectively. To improve semi-supervised multi-organ segmentation, DHC [31] proposes a dual-debiased heterogeneous co-training framework to enhance segmentation of difficult classes. SKCDF [40] decouples the information flow between labeled and unlabeled data by complementing their semantic knowledge. CGS [33] utilizes multiple segmentation heads to improve the model's perception of different organs.

2.2 Contrastive Learning in Semi-Supervised Medical Image Segmentation

Contrastive learning [9, 13] has shown great potential in enhancing feature representation. In SSMIS, methods like PH-Net [16] leverage contrastive learning to improve segmentation performance by aligning hard patches (identified by hardness scores) with high-confidence features in breast ultrasound images. Similarly, RD-Net [20] utilizes depth maps to enable the model to learn rich cross-modal complementary information for polyp segmentation. However, these methods are tailored to binary classification and may not generalize well to multi-organ scenarios. To address this, [3] proposed a pseudo-label guided contrastive loss to align the distributions of labeled and unlabeled data, while DyCON [1] introduced UnCL and FeCL to balance losses across organs. Despite these advancements, as the number of co-occurring organs increases, feature confusion between co-occurring pixels can lead to false activations. In this work, we focus on reducing noise in organ features and aim to mitigate the impact of co-occurrence in both the image and feature spaces.

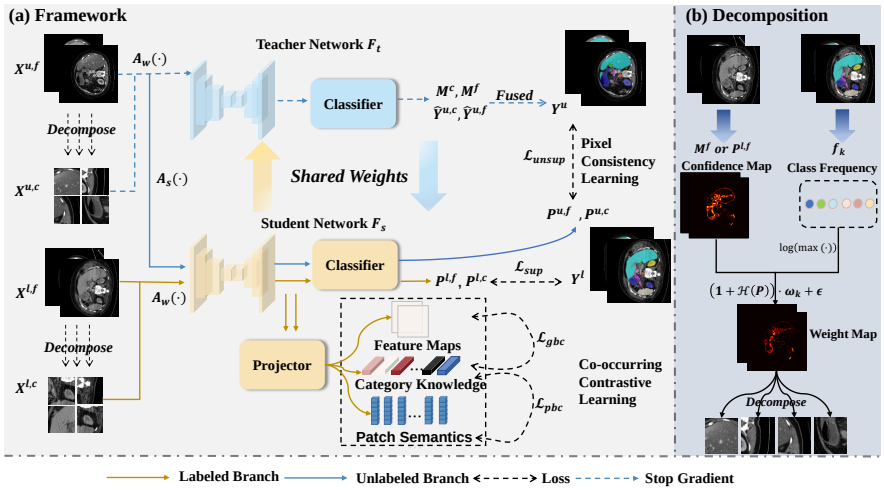


Fig. 2: The proposed **DPCL** framework and the decomposition process. (a) Overview of the proposed **DPCL** framework. The whole network consists of a labeled branch and an unlabeled branch. We design an *image decomposition mechanism* to enhance the model's perception of local regions, a *pseudo-label integration strategy* to generate high-quality pseudo labels, and a *co-occurring contrastive learning scheme* to mitigate the falsely activated regions caused by co-occurring organs. (b) Illustration of the decomposition process. We combine the confidence maps from the student and teacher networks with class frequency to construct a probabilistic sampling strategy for patch extraction.

3 Method

3.1 Preliminary

Given a medical image input $X \in \mathbb{R}^{W \times H \times L}$, the goal of semi-supervised medical image segmentation is to predict its pixel-wise semantic label $Y \in \{0, 1, \ldots, K-1\}^{W \times H}$, where K is the number of classes including the background. We denote the training dataset as $\mathcal{D} = \mathcal{D}^l \cup \mathcal{D}^u$, consisting of N labeled and M unlabeled full images ($N \ll M$). Specifically, $\mathcal{D}^l = \{(X_i^{l,f}, Y_i^{l,f})\}_{i=1}^{N} \cup \{(X_i^{l,c}, Y_i^{l,c})\}_{i=1}^{n \times N}$ and $\mathcal{D}^u = \{X_i^{u,f}\}_{i=1}^{M} \cup \{X_i^{u,c}\}_{i=1}^{n \times M}$, where $X^{l,f}$ and $X^{u,f}$ denote full labeled and unlabeled images, $Y^{l,f}$ is the ground-truth mask, and $X^{l,c}$, $X^{u,c}$, $Y^{l,c}$ are their cropped counterparts. Here n is the number of cropped patches generated per full image.

Our Net framework is shown in Fig. 2(a). This network consists of one teacher model F_t and one student model F_s. To accommodate both cropped and full image inputs, our network slightly deviates from the standard U-Net architecture. The detailed design is described in Sec. 3.2. Following [20], we define weak image augmentation (flipping, scaling and cropping) operations as A_w and strong image augmentation (color jitter, blurring, cutmix) operations as

A_s. By default, all images X have been subjected to A_w operations. Therefore, the probability maps are computed as:

$$\begin{aligned} P^{l,d} &= \sigma(F_s(X^{l,d})), \\ P^{u,d} &= \sigma(F_s(A_s(X^{u,d}))), \\ M^{d} &= \sigma(F_t(X^{u,d})), \end{aligned} \tag{1}$$

where $d \in \{c, f\}$ denotes the input type (*cropped* or *full*) and $\sigma(\cdot)$ denotes softmax function. The pseudo labels are obtained as:

$$\widehat{Y}^{u,d} = \arg\max(M^d) \tag{2}$$

Consequently, for labeled data, the overall supervised loss is defined as

$$\mathcal{L}_{sup} = \frac{1}{2|\mathcal{D}_l|} \sum_{i}^{|\mathcal{D}_l|} \sum_{d \in \{f,c\}} \mathcal{L}_{ce}(P_i^{l,d}, Y_i^{l,d}) + \mathcal{L}_{dice}(P_i^{l,d}, Y_i^{l,d}) \tag{3}$$

3.2 Local Enhancement

Image Decomposition. Owing to the limited effective regions in medical images, merely applying random cropping fails to enhance the model's sensitivity to minority classes. Therefore, as shown in Fig. 2(b), we adopt a class-aware sampling strategy to determine the cropping centers for both labeled images X^l and unlabeled images X^u, ensuring better coverage of rare categories and low-confidence regions.

Suppose the input is a full labeled image $X_i^{l,f}$, and its spatial domain is defined as $\Omega = (h, w)$. For each class k, the class frequency $f = [f_1, f_2, ..., f_{K-1}]$ represents the number of pixels belonging to class k in the ground-truth label $Y_i^{l,f}$. To emphasize minority classes, we define a logarithmic reweighting factor as

$$w_i^k = \log\big(\frac{\max(f) + 1}{f_k + 1}\big) \tag{4}$$

where $\max(\cdot)$ denotes the maximum value across all categories. We further incorporate a pixel-wise confidence map $P_i^{l,f}$ to balance class rarity and uncertainty. The final sampling weight for each pixel is defined as

$$W_{i,j}^{l,f} = \big(1 + \mathcal{H}(P_{i,j}^{l,f})\big) \cdot w_{Y_{i,j}^{l,f}} + \varepsilon, \tag{5}$$

where $\varepsilon > 0$ ensures non-zero weights and $j \in \Omega$. $Y_{i,j}^{l,f}$ and $P_{i,j}^{l,f}$ denote the values of $Y_i^{l,f}$ and $P_i^{l,f}$ at location j, respectively. The uncertainty is measured by the entropy $\mathcal{H}$, defined as $\mathcal{H}(P_{i,j}) = -\sum_{k=0}^{K-1} P_{i,j}(k) \log(P_{i,j}(k))$.

Based on the weight distribution $W^{l,f}$, we stochastically sample n centers to generate cropped regions. Given each sampled center and the predefined crop size s, we derive the crops $\{X_{im}^{l,c}\}_{m=1}^{n}$ and their corresponding spatial coordinates

$[\Omega_{i1}, \ldots, \Omega_{in}]$. For unlabeled $X_i^{u,f}$, this process uses $\widehat{Y}_i^{u,f}$ and M_i^f instead.

Dual Normalization. Due to the domain gap between cropped and full images (see in Fig. 1, the model not only fails to capture fine-grained details but also suffers from domain conflicts, leading to degraded segmentation performance. Inspired by [6], to narrow the domain gap between cropped images and full images efficiently, we design dual-normalization U-Net. Specifically, the proposed dual-normalization (DN) module can be formulated as

$$\mathrm{DN}(z; d \in \{c, f\}) = \gamma_d \frac{z - \mu_d}{\sqrt{\sigma_d^2 + \varepsilon}} + \beta_d, \tag{6}$$

where z denotes the activated features extracted from cropped or full images. During training, the DN module estimates means and variances by exponential moving average with update factor β.It can be written as

$$\bar{\mu}_d^{t+1} = (1 - \beta)\bar{\mu}_d^t + \beta \mu_d^t \tag{7}$$

$$(\bar{\sigma}_d^{t+1})^2 = (1 - \beta)(\bar{\sigma}_d^t)^2 + \beta(\sigma_d^t)^2 \tag{8}$$

where t denotes the current training iteration, $\bar{\mu}_d^t$ and $\left(\bar{\sigma}_d^t\right)^2$ are the estimated means and variances of d in the t-th iteration.

Pixel Consistency For unlabeled images $X_i^{u,f}$, the quality of pseudo labels plays a pivotal role in improving segmentation accuracy and overall model consistency. Since each $X_i^{u,f}$ yields n cropped patches, there are n corresponding probability maps $\{M_{im}^c\}_{m=1}^n$ and pseudo labels $\{\widehat{Y}_{im}^{u,c}\}_{m=1}^n$. The probability maps are then aggregated into a unified map as

$$\widetilde{M}_{i,j}^c = \begin{cases} M_{im,j}^c, & \text{if } j \in \Omega_{im} \text{ and } j \notin \mathcal{R}, \\ M_{i\hat{m},j}^c, & \text{if } j \in \mathcal{R}, \\ \infty, & \text{otherwise}, \end{cases} \tag{9}$$

where $\mathcal{R}$ denotes the set of overlapping pixels, and $\hat{m} = \arg\min_m \mathcal{H}(M_{i,m,j}^c)$ selects the patch with the lowest entropy at pixel j. Following the same strategy, the pseudo labels are merged to form $\widetilde{Y}_i^{u,c}$. To obtain more reliable pseudo labels, we leverage the complementary segmentation capabilities of the model on both cropped and full images to generate a more consistent and robust supervision signal:

$$O_{i,j} = \begin{cases} \mathbf{1}\left[\mathcal{H}(\widetilde{M}_{i,j}^c) < \mathcal{H}(M_{i,j}^f)\right], & \text{if } \widetilde{M}_{i,j}^c \neq \infty \\ 1, & \text{if } \widetilde{M}_{i,j}^c = \infty \end{cases} \tag{10}$$

Accordingly, the refined pseudo labels and probability maps can be formulated as:

$$\begin{aligned} Y_i^u &= O_i \odot \widetilde{Y}_i^{u,c} + (1 - O_i) \odot \widehat{Y}_i^{u,f}, \\ M_i^u &= O_i \odot \max(\widetilde{M}_i^c) + (1 - O_i) \odot \max(M_i^f). \end{aligned} \tag{11}$$

Therefore, the reliable pseudo label Y^u can be used to supervise the predictions of both the cropped and full images, which can be formulated as follows:

$$\mathcal{L}_{unsup} = \frac{1}{|\mathcal{D}_u|} \sum_{i}^{|\mathcal{D}_u|} \sum_{d \in \{f,c\}} \mathcal{L}_{dice}(P_i^{u,d}, Y_i^u) \cdot A_i \tag{12}$$

where

$$A_i = \mathbf{1}[M_i^u \geq \gamma_1] \tag{13}$$

where γ_1 is a pre-defined confidence threshold to filter out unreliable pixels.

3.3 Representation with Co-occurring Contrastive Learning

Previous semi-supervised segmentation methods [16, 20] typically perform contrastive learning by filtering out unreliable features. However, they are primarily designed for binary segmentation tasks. In multi-organ scenarios, as shown in Fig. 5, previous methods often activate co-occurring regions with high confidence, limiting their ability to suppress false activations and disentangle correlated structures. To overcome this limitation, we propose a co-occurring contrastive learning strategy built upon the patch decomposition process introduced in Sec. 3.2.

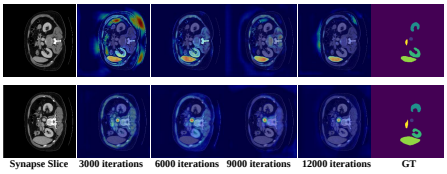


Fig. 3: Grad-CAM visualization on Synapse dataset during training.

Dual Domain Class-Specific Prototypes. To reduce noise in prototype construction, we compute class prototypes using only labeled images and filter out unreliable features. For the input $X_i^{l,f}$ or $X_i^{l,c}$, the prototype for class k is calculated by averaging the reliable features belonging to that class:

$$\mathcal{G}_i^{k,d} = \frac{\sum_{j \in \Omega} g_{i,j}^d \, \mathbf{1}[Y_{i,j}^{l,d} = k] \, \mathbf{1}[P_{i,j}^{l,d}(k) \geq \gamma_2]}{\sum_{j \in \Omega} \mathbf{1}[Y_{i,j}^{l,d} = k] \, \mathbf{1}[P_{i,j}^{l,d}(k) \geq \gamma_2]}, \tag{14}$$

where $g_{i,j}^d$ denotes the projected feature of the j-th pixel from the student model, and γ_2 denotes the confidence threshold (default: 0.9) for filtering unreliable features.

Updating Prototypes. In the prototype-based contrastive learning for multi-class segmentation, classes with fewer occurrences are often mixed with noise due to slower updates of their prototypes. To address this, we leverage features obtained from cropped images to accelerate the updating of prototypes. Specifically, image patches obtained in Sec. 3.2, which contain small classes, can be used to accelerate the update of prototypes for these underrepresented classes. The updating process is formulated as:

$$\mathcal{G}_k^{t+1} = \eta \mathcal{G}_k^t + \frac{(1-\eta)}{2} \mathcal{G}_i^{k,c} + \frac{(1-\eta)}{2} \mathcal{G}_i^{k,f} \tag{15}$$

where η is a hyperparameter, empirically set to 0.99 in the experiments, and $\mathcal{G}_k^t$ denotes the prototypes after t iterations.

Patch-based Contrastive Loss. To separate the features of co-occurring classes, we first construct patch semantics $q_i^c = [q_{i0}, q_{i1}, \ldots, q_{i,K-1}]$ for i-th cropped image:

$$q_{ik} = \frac{\sum_{j\in\Omega} g_{i,j}^c \mathbf{1}[Y_{i,j}^{l,c} = k]}{\sum_{j\in\Omega} \mathbf{1}[Y_{i,j}^{l,c} = k]}, \tag{16}$$

With patch semantics and global prototypes, we construct the $\mathcal{L}_{pbc}$ loss, inspired by InfoNCE [25], to guide the student network. To reduce false activations caused by feature confusion among co-occurring organs, we focus on decoupling the features of neighboring organs within local patches. Specifically, only the features from co-occurring classes in cropped images are used as negative samples, while the corresponding global prototype $\mathcal{G}_k^t$ of the same class serves as the positive sample. This strategy strengthens the separation of co-occurring organs, ensuring more accurate segmentation results. The loss can be formulated as:

$$\mathcal{L}_{pbc} = -\frac{1}{N_p^+} \sum_{i=1}^{n\times N} \sum_{k=0}^{K-1} \log \frac{\exp(q_{ik}^T \mathcal{G}_k^t/\tau)}{\sum_{m=0}^{K_c-1} \exp(q_{ik}^T \mathcal{G}_m^t/\tau)}, \tag{17}$$

where τ is the temperature factor, K_c denotes the number of co-occurring classes within the cropped image, N_p^+ counts the number of positive pairs, and $^\top$ indicates the transpose operation.

Global-based Contrastive Loss. While $\mathcal{L}_{pbc}$ effectively decouples features between co-occurring classes within local patches, relying solely on this loss may lead to feature overlap and confusion among non-co-occurring organs, as shown in Fig. 6a. To ensure the model benefits from both the broader field of view in full images and the fine-grained details in cropped images, we introduce the global-based contrastive loss. Specifically, we treat the prototypes of the same class as positive samples and those from different classes as negative samples. The global contrastive loss is defined as:

$$\mathcal{L}_{gbc} = -\frac{1}{N_g^+} \sum_{i=1}^{N} \sum_{j\in\Omega} \log \frac{\exp\left((g_{i,j}^f)^T \mathcal{G}_{Y_{i,j}^{l,f}}^t/\tau\right)}{\sum_{k=0}^{K-1} \exp\left((g_{i,j}^f)^T \mathcal{G}_k^t/\tau\right)}, \tag{18}$$

where $\mathcal{G}_{Y_{i,j}^{l,f}}^t$ denotes the prototype of the class corresponding to the ground-truth label $Y_i^{l,f}$ at position j. As shown in Fig. 3, the decoupling process enables the model to gradually concentrate on the target regions with increasing iterations.

3.4 Traning Objectives

As shown in Fig. 2(a), our model is optimized using a combination of supervised segmentation loss $\mathcal{L}_{sup}$, unsupervised segmentation loss $\mathcal{L}_{unsup}$, and two

Table 1: Results of Synapse dataset with 5% and 10% annotations. "L/U" indicates the labeled and unlabeled samples.

L / U	Method	Venue	Average Dice (%) ↑	Average ASD ↓	Average Dice of Each Class (%) ↑							
					Ao.	Ga.	LK.	RK.	Li.	Pa.	Sp.	St.
1 / 17 (5 / 95%)	Sup-Only	-	32.98	61.59	50.24	19.29	24.52	32.63	77.45	9.7	41.57	8.47
	UA-MT [39]	MICCAI'19	34.36	45.59	64.42	21.88	26.46	29.91	73.31	11.75	34.22	12.89
	FixMatch [29]	NIPS'20	48.83	33.57	76.53	38.78	49.69	46.66	85.29	6.68	66.15	20.85
	SS-Net [34]	MICCAI'22	33.81	49.50	65.73	8.21	39.16	37.58	72.73	9.35	33.52	6.22
	BCP [2]	CVPR'23	42.88	53.51	73.31	17.41	50.29	55.99	65.12	6.02	44.00	30.88
	ABD [10]	CVPR'24	53.57	32.84	72.78	17.46	59.04	60.66	81.53	26.13	78.79	34.18
	AD-MT [42]	ECCV'24	47.08	35.76	68.25	**43.72**	19.47	33.79	84.57	22.74	71.33	32.79
	CGS [33]	TMI'25	49.55	46.57	80.99	21.94	**74.89**	**71.78**	81.94	9.11	45.07	10.65
	M3HL [21]	MICCAI'25	33.19	40.67	73.00	10.84	59.34	51.38	27.87	9.73	29.47	3.88
	SKCDF [40]	CVPR'25	29.52	24.28	44.45	19.45	20.13	30.71	80.91	3.34	27.50	9.65
	Ours	This paper	**61.48**	**23.78**	**87.62**	31.98	60.64	63.94	**87.77**	**41.26**	**84.08**	**35.10**
2 / 16 (10 / 90%)	Sup-Only	-	37.75	43.54	58.06	23.07	39.33	27.94	81.00	10.48	40.71	21.39
	UA-MT [39]	MICCAI'19	39.94	42.24	69.96	26.67	44.14	27.30	83.46	4.36	44.72	18.90
	FixMatch [29]	NIPS'20	60.48	16.66	78.93	31.62	67.87	66.30	89.57	36.94	70.98	41.61
	SS-Net [34]	MICCAI'22	39.57	49.42	70.85	15.53	23.19	42.71	78.80	17.06	49.27	19.17
	BCP [2]	CVPR'23	56.83	36.78	76.94	17.73	68.74	68.36	86.19	34.70	66.16	35.84
	ABD [10]	CVPR'24	62.55	23.17	82.40	15.15	**77.68**	**76.58**	87.07	31.34	83.60	23.17
	AD-MT [42]	ECCV'24	62.73	22.25	63.34	**42.91**	71.01	73.23	**91.35**	32.35	74.75	**52.93**
	CGS [33]	TMI'25	59.20	27.81	83.17	25.29	62.38	66.44	88.28	38.83	78.43	30.78
	M3HL [21]	MICCAI'25	40.06	33.18	73.74	33.29	65.43	47.40	29.07	15.27	37.21	19.05
	SKCDF [40]	CVPR'25	51.09	**12.38**	74.40	19.01	62.09	65.38	88.48	11.68	72.24	15.45
	Ours	This paper	**68.89**	14.84	**84.36**	39.62	74.49	71.65	91.21	**54.90**	**85.77**	48.92

contrastive learning losses $\mathcal{L}_{pbc}$ and $\mathcal{L}_{gbc}$. The overall optimization objective is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_u \mathcal{L}_{unsup} + \lambda_m \big(\mathcal{L}_{pbc} + \mathcal{L}_{gbc}\big) \tag{19}$$

where λ_u and λ_m are loss weights to balance the relationship between losses (specifically, following [14], we choose $\lambda_u = 0.25$ and λ_m as Gaussian warming up function $\lambda(t) = 0.1 \times e^{-5(1-t/t_{max})^2}$, where t_{max} is the maximum training step)

4 Experiments and Results

4.1 Experiments setup

Datasets. We evaluated our method on four pubic medical image datasets:

- **Synapse dataset** [18] contains 30 abdominal CT scans and there are 3779 axial contrast-enhanced abdominal CT images in total with 9 distinct classes. Strictly following [33], 18 cases are divided for training and the other 12 cases are for testing.
- **FeTA2021 dataset** [26] contains 80 T2-weighted fetal brain MRI images, manually annotated into 8 distinct classes. We split the dataset into 56 training, 8 validation, and 16 test images at a ratio of 7:1:2.

Table 2: Comparison of Dice and ASD Metrics Among Different Semi-Supervised Methods on FeTA2021, SegTHOR, and ACDC Datasets. In M3HL, the values in () represent the reported data, while the outer values are the reproduced data.

Method	Feta2021				SegTHOR				ACDC			
	5%		10%		10%		20%		5%		10%	
	Dice(%)↑	ASD↓	Dice(%)↑	ASD↓	Dice(%)↑	ASD↓	Dice(%)↑	ASD↓	Dice(%)↑	ASD↓	Dice(%)↑	ASD↓
Sup-Only	52.26	8.80	69.60	5.32	64.98	18.34	73.13	10.35	47.83	12.62	79.41	2.70
FixMatch [29]	57.29	5.89	72.37	5.51	76.65	4.20	77.91	3.88	87.27	1.15	88.89	1.37
SS-Net [34]	57.92	5.95	74.65	2.87	70.27	6.10	75.79	4.14	46.04	7.75	81.65	2.02
BCP [2]	64.33	7.27	75.15	4.34	75.32	3.86	76.59	3.37	87.59	0.67	88.84	1.17
ABD [10]	59.93	15.18	76.12	3.64	73.46	4.68	77.56	3.72	88.92	0.52	89.81	0.49
AD-MT [42]	68.31	4.60	76.08	3.69	75.89	5.38	78.26	5.49	88.75	0.50	89.46	0.44
CGS [33]	64.58	7.04	75.99	5.13	73.69	4.48	76.87	4.31	88.83	0.66	89.83	0.68
M3HL [21]	50.77	7.66	72.67	3.22	72.26	6.89	80.08	3.84	87.44	0.59	88.83(90.47)	0.74(0.34)
Ours	**70.04**	**4.58**	**77.10**	**2.29**	**78.18**	**3.26**	**81.38**	**3.15**	**89.28**	**0.34**	**90.47**	**0.28**

– **SegTHOR dataset** [17] consists of 40 CT volumes. The goal of the SegTHOR challenge is to automatically segment 4 organs-at-risk (OARs): the heart, aorta, trachea, and esophagus. Following [33], we split the dataset into 28 cases for training, 4 for validation, and 8 for testing.
– **ACDC dataset** [4] consists of 200 annotated short-axis cardiac MR images from 100 patients across 4 classes. Following [2], we split the dataset into 70 cases for training, 10 for validation, and 20 for testing.

For all datasets, we convert the 3D volumes into 2D slices and resize them to 256×256. We adopt two widely used evaluation metrics, including the Dice Score (%) and Average Surface Distance (ASD) in voxels. Dice measures the overlap ratio between the prediction and the ground truth, while ASD computes the average boundary distance between them.

4.2 Implementation Details

For fair comparison, following most of the previous work [2, 10, 33, 42], we use U-Net [27] as the backbone. All normalization layers in the model are replaced with Dual Normalization. The projector outputs features with a dimension of 256, and the number of cropped images is set to 4 by default, with the crop size set to 64. The student model is trained using a stochastic gradient descent (SGD) optimizer with 0.9 momentum and a weight decay of 0.0001. The initial learning rate is set to 0.01, which is decayed using a polynomial strategy: $lr = lr_{\text{init}} \cdot \left(1 - \frac{t}{t_{\max}}\right)^{0.9}$. For experimental fairness, we set the batch size to 6 for both labeled and unlabeled data, and train for 30,000 iterations. Notably, during inference, we use only the full images. All experiments are conducted on an NVIDIA 3090 GPU.

4.3 Comparison with State-of-the-Art Methods

We compare our method with several state-of-the-art semi-supervised segmentation techniques [2, 10, 21, 29, 33, 34, 39, 40, 42]. All methods are evaluated using

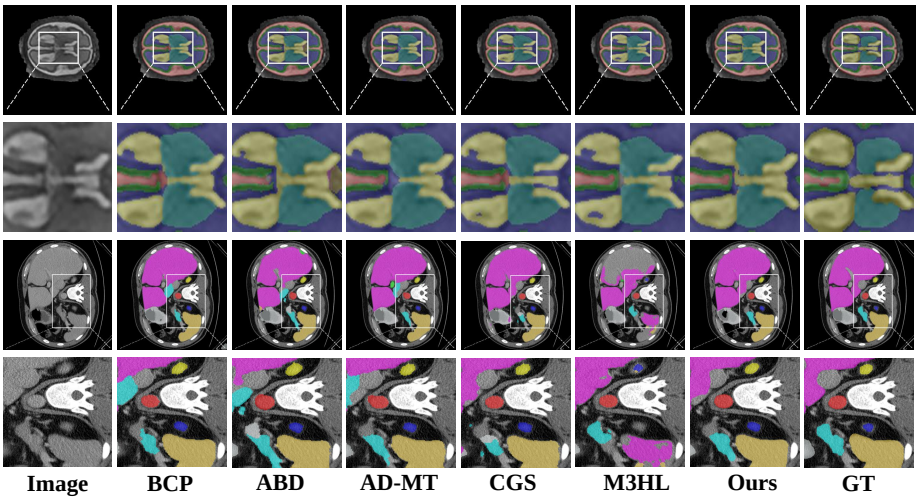


Fig. 4: Visual comparison of different state-of-the-art methods on the 10% labeled Synapse dataset and 20% labeled FeTA dataset. Ground truth labels are denoted as GT.

the same labeled and unlabeled datasets for a fair comparison.

Synapse Dataset. As presented in Tab. 1, we report the segmentation results for the Synapse dataset under 5% and 10% labeled data settings. Our method consistently outperforms all other SOTA approaches, achieving a substantial Dice score improvement of 22.26% and 9.78% in the 5% and 10% cases, respectively. It is noteworthy that certain organs, such as the pancreas and stomach, are particularly challenging to segment when labeled data is scarce. Under the extremely limited 5% labeled condition, our method significantly boosts the Dice score for the pancreas and stomach by 57.90% and 2.9%, respectively, demonstrating its robust capability in handling difficult organ structures.

FeTA2021 Dataset. In Tab. 2, we further evaluate the performance of various methods on the FeTA2021 dataset using only 3 labeled samples (5%) or 6 labeled samples (10%). Our approach achieves a Dice score improvement of 2.53% and 1.29% for the 5% and 10% labeled scenarios, respectively, surpassing other competing methods and showing superior performance even with a very small number of labeled samples.

SegTHOR Dataset. We also validate our method on the SegTHOR dataset using 3 labeled samples (10%) and 6 labeled samples (20%), as shown in Tab. 2. Our method maintains a consistent lead over existing techniques, yielding Dice score improvements of 1.99% and 1.62% for the 10% and 20% labeled data settings, respectively. This confirms the stability and effectiveness of our framework across different labeling ratios.

ACDC Dataset. Finally, Tab. 2 provides the comparative results on the ACDC dataset with 3 labeled samples (5%) and 7 labeled samples (10%). Again, our

Table 3: Ablation experiments on 10% labeled Synapse dataset.

Sup-Only	Local Enhance.	PBC	GBC	DN	Dice(%) ↑ Ao.	Ga.	LK.	RK.	Li.	Pa.	Sp.	St.	Dice(%) ↑ Avg.	ASD ↓ Avg.
✓					58.09	23.07	39.33	27.94	81.00	10.48	40.71	21.39	37.75	43.54
	✓			✓	83.40	36.25	69.78	71.31	89.62	44.23	80.90	40.94	64.55	20.85
	✓	✓		✓	85.83	34.82	73.30	71.10	90.02	47.13	81.70	45.48	66.17	16.11
	✓		✓	✓	**86.92**	36.17	**75.03**	71.18	**91.28**	48.17	84.61	46.69	67.50	15.88
	✓	✓	✓		83.15	34.42	71.28	71.06	89.50	44.37	80.64	41.32	64.84	20.48
	✓	✓	✓	✓	84.36	**39.62**	74.49	**71.65**	91.21	**54.90**	**85.77**	**48.92**	**68.89**	**14.84**

Table 4: Ablation studies on hyperparameters and decomposition strategies. The best results are highlighted in **bold**.

(a) Effect of n

n	Dice (%) ↑	ASD ↓
2	67.91	15.62
4	**68.89**	**14.84**
6	68.53	15.21
8	67.34	16.40

(b) Effect of s

s	Dice (%) ↑	ASD ↓
16	66.23	18.46
32	67.78	15.61
64	**68.89**	**14.84**
128	67.27	17.12

(c) Decomposition strategies

Strategy	Dice (%) ↑	ASD ↓
Uniform	65.57	20.26
Uncertainty	67.54	15.61
Class Freq.	66.10	16.27
Ours	**68.89**	**14.84**

proposed method achieves superior results, outperforming other state-of-the-art methods across all evaluated metrics, further establishing its generalizability and effectiveness in semi-supervised medical image segmentation.

Furthermore, we present the visual results for these four datasets in Fig. 4, and the visualization of CAMs for Synapse in Fig. 5. The comparison shows that our method outperforms others in semi-supervised multi-organ segmentation.

4.4 Ablation studies

We conduct a series of ablation studies to validate the effectiveness of each module in our framework. All experiments below are performed on the Synapse dataset using 10% labeled data.

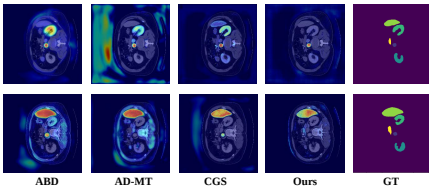
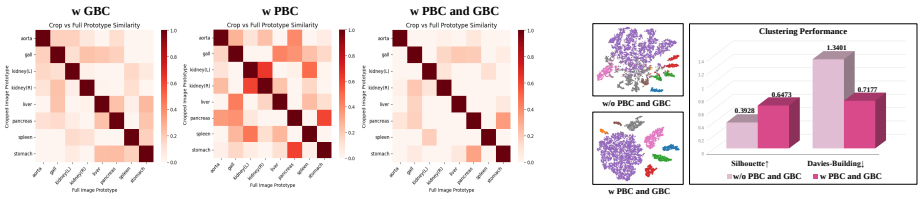


Fig. 5: Visualization of CAMs on the 10% labeled Synapse dataset.

Effectiveness of Each Module. Tab. 3 presents a stepwise ablation study on each module. We set the model trained with only supervised learning (Sup-Only) as our baseline, which only achieves a Dice score of 37.75%. After incorporating Local Enhancement, the model's Dice score increases by 6%, demonstrating the module's ability to provide effective local information and enhance segmentation performance in local regions. Building upon this, adding the patch-based contrastive loss (PBC) further improves the Dice score by 3.39%, while incorporating the global-based contrastive loss (GBC) results in a 5.73%



(a) Prototype heatmaps **(b)** t-SNE and clustering metrics

Fig. 6: Visualization of feature learning. (a) Heatmaps of prototypes under different losses. (b) t-SNE visualization of feature space with clustering metrics.

improvement. The complete model achieves a Dice score of 68.67%; if the Dual Normalization module is absent, the Dice score drops by 5.83%, which strongly validates the effectiveness of our approach for multi-organ segmentation.

Image Decomposition Strategies.

As shown in Tab. 4, to validate the effectiveness of our image decomposition strategy, we compare our approach with three other sampling weight methods: uniform distribution, uncertainty-based, and class frequency-based. The uniform distribution method, which does not incorporate any additional information, results in a Dice score of only 65.57%, failing to provide meaningful local information to the model. While the uncertainty based and class frequency based methods both achieve improvements over the uniform method, with respective Dice score increases of 0.8% and 3.0%, neither reaches the performance attained by the fusion of both strategies. This highlights the advantage of combining these strategies for more effective local enhancement.

Selection of Cropping Parameters. We analyze the effect of the number of cropped patches n, crop size s. As shown in Tab. 4, the model achieves the best performance in terms of both Dice and ASD when $n = 4$ and $s = 64$. Ablation studies on other hyperparameters can be found in the Appendix.

Feature Representation Analysis. We further visualize the feature representations to validate the effectiveness of our method. Specifically, we compare the heatmaps of prototypes derived from cropped and full images under different losses. As shown in Fig. 6a, when only the GBC loss is applied, adjacent organs such as the pancreas and liver exhibit poor separability. In contrast, incorporating PBC alleviates this issue by enhancing the discriminability of neighboring regions. Furthermore, we employ t-SNE [24] to project the feature embeddings into 2D space. After integrating both PBC and GBC, features within the same class become more compactly clustered, leading to a 3.44% improvement in the Silhouette Score [28] and a 5.34% reduction in the Davies–Bouldin Index [11].

5 Conclusion

In this paper, we introduced a novel semi-supervised medical image segmentation framework that effectively addresses the feature confusion caused by organ

co-occurrence, which becomes more severe under limited annotations. Specifically, we designed a local enhancement strategy to improve the model's perception of local regions and a decoupling mechanism for co-occurrence organs to achieve more discriminative feature representations. Extensive experiments on four benchmark multi-organ datasets demonstrate that our method consistently outperforms existing state-of-the-art semi-supervised methods, highlighting its effectiveness and strong generalization capability.

Acknowledgements

This research is supported by National Natural Science Foundation of China No. 62371156 and Shenzhen Science and Technology Program No. JCYJ20240813105132043.

References

1. Assefa, M., Naseer, M., Ganapathi, I.I., Ali, S.S., Seghier, M.L., Werghi, N.: Dycon: Dynamic uncertainty-aware consistency and contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30850–30860 (2025) 2, 4
2. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 11514–11524 (2023) 2, 4, 10, 11
3. Basak, H., Yin, Z.: Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19786–19797 (2023) 4
4. 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 transactions on medical imaging **37**(11), 2514–2525 (2018) 11
5. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Medical Image Analysis **87**, 102792 (2023) 2
6. Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 7354–7362 (2019) 7
7. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021) 2
8. Chen, L., Lei, C., Li, R., Li, S., Zhang, Z., Zhang, L.: Fpr: False positive rectification for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1108–1118 (2023) 3
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020) 4
10. Chi, H., Pang, J., Zhang, B., Liu, W.: Adaptive bidirectional displacement for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4070–4080 (2024) 2, 4, 10, 11

11. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence (2), 224–227 (2009) 14
12. Dou, Q., Liu, Q., Heng, P.A., Glocker, B.: Unpaired multi-modal segmentation via knowledge distillation. IEEE transactions on medical imaging **39**(7), 2415–2425 (2020) 2
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 4
14. Huang, H., Huang, Y., Xie, S., Lin, L., Tong, R., Chen, Y.W., Li, Y., Zheng, Y.: Combinatorial cnn-transformer learning with manifold constraints for semi-supervised medical image segmentation. In: AAAI Conference on Artificial Intelligence. vol. 38, pp. 2330–2338 (2024) 4, 10
15. Huang, W., Zhang, L., Wang, Z., Wang, L.: Exploring inherent consistency for semi-supervised anatomical structure segmentation in medical imaging. IEEE Transactions on Medical Imaging (2024) 4
16. Jiang, S., Wu, H., Chen, J., Zhang, Q., Qin, J.: Ph-net: semi-supervised breast lesion segmentation via patch-wise hardness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11418–11427 (2024) 4, 8
17. Lambert, Z., Petitjean, C., Dubray, B., Kuan, S.: Segthor: Segmentation of thoracic organs at risk in ct images. In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6. Ieee (2020) 11
18. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015) 10
19. Lei, M., Wu, H., Lv, X., Wang, X.: Condseg: A general medical image segmentation framework via contrast-driven feature enhancement. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4571–4579 (2025) 2
20. Li, Y., Zhu, Z., Zhang, Y., Chen, Y., Yu, Z.: Boost the inference with co-training: A depth-guided mutual learning framework for semi-supervised medical polyp segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10394–10403 (2025) 4, 5, 8
21. Liu, Y., Zhang, Z., Yue, J., Guo, W., Li, D.: M 3 hl: Mutual mask mix with high-low level feature consistency for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 314–323. Springer (2025) 10, 11
22. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 4
23. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: International Conference on Medical Imaging with Deep Learning. pp. 820–833. PMLR (2022) 4
24. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008) 14
25. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 9

26. Payette, K., de Dumast, P., Kebiri, H., Ezhov, I., Paetzold, J.C., Shit, S., Iqbal, A., Khan, R., Kottke, R., Grehten, P., et al.: An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. Scientific data **8**(1), 167 (2021) 10
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 234–241. Springer (2015) 11
28. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987) 14
29. 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. Neural Information Processing Systems **33**, 596–608 (2020) 10, 11
30. Su, J., Luo, Z., Lian, S., Lin, D., Li, S.: Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. Medical Image Analysis **94**, 103111 (2024) 2
31. Wang, H., Li, X.: Dhc: Dual-debiased heterogeneous co-training framework for class-imbalanced semi-supervised medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 582–591. Springer (2023) 4
32. Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L.: Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Medical image analysis **55**, 88–102 (2019) 2
33. Wang, Y., Li, Z., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Balancing multi-target semi-supervised medical image segmentation with collaborative generalist and specialists. IEEE Transactions on Medical Imaging (2025) 2, 4, 10, 11
34. Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J.: Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 34–43. Springer (2022) 10, 11
35. Xia, K., Wang, L., Zhou, S., Zheng, N., Tang, W.: Learning to refactor action and co-occurrence features for temporal action localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13884–13893 (2022) 3
36. Xu, W., Li, C., Bian, Y., Meng, Q., Zhu, W., Shi, F., Chen, X., Shao, C., Xiang, D.: Cross-modal consistency for single-modal mr image segmentation. IEEE Transactions on Biomedical Engineering **71**(9), 2557–2567 (2024) 2
37. Yang, Z., Fu, K., Duan, M., Qu, L., Wang, S., Song, Z.: Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3606–3615 (2024) 3
38. You, C., Zhou, Y., Zhao, R., Staib, L., Duncan, J.S.: Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging **41**(9), 2228–2237 (2022) 2
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: Medical Image Computing and Computer Assisted Intervention. pp. 605–613. Springer (2019) 2, 4, 10, 11
40. Zhang, Z., Yin, G., Zhang, B., Liu, W., Zhou, X., Wang, W.: A semantic knowledge complementarity based decoupling framework for semi-supervised class-imbalanced medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25940–25949 (2025) 2, 4, 10, 11

41. Zhao, H., Meng, H., Yang, D., Xie, X., Wu, X., Li, Q., Niu, J.: Guidednet: Semi-supervised multi-organ segmentation via labeled data guide unlabeled data. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 886–895 (2024) 2
42. Zhao, Z., Wang, Z., Wang, L., Yu, D., Yuan, Y., Zhou, L.: Alternate diverse teaching for semi-supervised medical image segmentation. In: European Conference on Computer Vision. pp. 227–243. Springer (2024) 4, 10, 11