International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02328-9

PICK: Predict and Mask for Semi-supervised Medical Image Segmentation

Qingjie Zeng[1] · Zilin Lu[1] · Yutong Xie[2] · Yong Xia[1,3,4]

Received: 2 July 2024 / Accepted: 9 December 2024

Abstract

Pseudo-labeling and consistency-based co-training are established paradigms in semi-supervised learning. Pseudo-labeling focuses on selecting reliable pseudo-labels, while co-training emphasizes sub-network diversity for complementary information extraction. However, both paradigms struggle with the inevitable erroneous predictions from unlabeled data, which poses a risk to task-specific decoders and ultimately impact model performance. To address this challenge, we propose a PredICt-and-masK (PICK) model for semi-supervised medical image segmentation. PICK operates by masking and predicting pseudo-label-guided attentive regions to exploit unlabeled data. It features a shared encoder and three task-specific decoders. Specifically, PICK employs a primary decoder supervised solely by labeled data to generate pseudo-labels, identifying potential targets in unlabeled data. The model then masks these regions and reconstructs them using a masked image modeling (MIM) decoder, optimizing through a reconstruction task. To reconcile segmentation and reconstruction, an auxiliary decoder is further developed to learn from the reconstructed images, whose predictions are constrained by the primary decoder. We evaluate PICK on five medical benchmarks, including single organ/tumor segmentation, multi-organ segmentation, and domain-generalized tasks. Our results indicate that PICK outperforms state-of-the-art methods. The code is available at https://github.com/maxwell0027/PICK.

Keywords Semi-supervised learning · Medical image segmentation · Attentive region masking · Reconstruction

1 Introduction

Accurate segmentation of medical images is crucial in clinical practice, as the segmented regions provide vital insights into organ/tumor characteristics such as volume, location, and shape (Antonelli et al., 2022). Recent studies (Kirillov et al., 2023; Li et al., 2022; Wang & Li, 2023) highlight the significant potential of data-driven deep learning models, which demonstrate superior performance across various vision and language tasks (Liu et al., 2021; Agrawal et al., 2019; Xie et al., 2022). However, achieving promising results with these models requires a large amount of paired image-label data. Unfortunately, in medical image segmentation, obtaining annotations for volumetric data is challenging due to the specialized expertise required from clinical physicians Xie et al. (2023). In scenarios with predominantly unlabeled data, semi-supervised learning (SSL) has emerged as a popular paradigm for medical image segmentation (Bai et al., 2023; Zeng et al., 2023; Chen et al., 2023), leveraging both limited labeled and abundant unlabeled data efficiently.

Communicated by Ziyue Xu.

✉ Yutong Xie
yutong.xie678@gmail.com

✉ Yong Xia
yxia@nwpu.edu.cn

Qingjie Zeng
qjzeng@mail.nwpu.edu.cn

Zilin Lu
luzl@mail.nwpu.edu.cn

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Northwestern Polytechnical University, 1 Dongxiang Road, Xi'an 710072, Shaanxi, China

[2] Australian Insititute for Machine Learning, The University of Adelaide, Adelaide, SA 5000, Australia

[3] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China

[4] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China

Published online: 05 January 2025

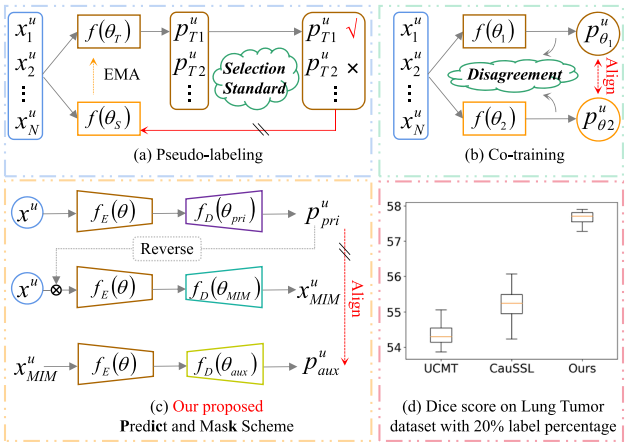

Fig. 1 Comparison of our proposed PICK with existing popular semi-supervised schemes. **a** Pseudo-labeling emphasizes the generation of reliable pseudo-labels, which typically involves developing suitable selection criteria and collaborating with a teacher model that is updated using an exponential moving average (EMA) of the student's parameters. **b** Co-training aims to achieve consistency while addressing discrepancies. These methods generally employ diverse backbone architectures with different parameter initializations and utilize specialized schemes to ensure that the models focus on complementary regions. Both **a** and **b** introduce unavoidable incorrect predictions that can adversely affect the decision-making processes of task-specific decoders. In contrast, **c** our PICK introduces a novel predict-and-mask strategy for learning from unlabeled data. This approach distinguishes itself from existing methods by employing a pseudo-label-guided reconstruction task to extract information from unlabeled data, enabling the transfer of unlabeled target information into the encoder while preserving the integrity of the primary decoder. **d** The performance comparison illustrates that our PICK significantly outperforms existing methods

Current SSL approaches primarily fall into two categories: pseudo-labeling (Ma et al., 2023; Zeng et al., 2023) and consistency-based co-training. Pseudo-labeling methods aim to select high-quality pseudo-labels by establishing reliable criteria for label selection (see Fig. 1a), such as using adaptive thresholds (Ma et al., 2023) or prior loss distributions (Zeng et al., 2023). Another approach involves ensembling multiple teachers to supervise a student's learning progress, akin to a mean-teacher strategy (Bai et al., 2023). However, these strategies often encounter issues of over-confidence or confirmation bias, where incorrect predictions with high classification scores can mislead the learning process (Guo et al., 2017; Jin et al., 2022). On the other hand, consistency-based co-training emphasizes increasing disagreement between sub-networks, typically achieved by leveraging sub-networks with different architectures or initialization weights in a cross-pseudo supervision scheme (see Fig. 1b). Despite its potential benefits, this approach faces a consensus problem where sub-networks tend to converge quickly in the early training stages, potentially leading to the degradation of co-training into a vanilla self-training method (Shen et al., 2023). To address this challenge, researchers have explored alternative teacher designs (Na et al., 2023) and uncertainty estimation schemes (Peiris et al., 2023) to regulate sub-network focus on diverse regions, thereby promoting consistency in learning while managing discrepancies (Zeng et al., 2023). However, balancing stability and discrepancy remains a persistent challenge in co-training due to inherent conflicts between these two objectives. Furthermore, effective implementation of cross-pseudo supervision in co-training necessitates reliable and complementary pseudo-labels, a combination that remains difficult to achieve simultaneously.

Faced with the aforementioned dilemma, we intuitively consider: *given that erroneous boundaries in pseudo-labels are inevitable, yet the target regions they indicate are generally correct, is there a method to transfer this target region information from pseudo-labels to the encoder during training, rather than allowing pseudo-labels to directly influence the decoder?* Achieving this could mitigate the direct risk of erroneous predictions by the decoder and alleviate the demanding requirements of co-training. In segmentation tasks, the encoder is responsible for extracting high-level features from input images, capturing essential spatial and semantic information. The decoder then utilizes these features to produce a pixel-wise segmentation map, translating abstract representations back into the image space. This clear division of labor ensures that the decoder remains focused on the task, as it is directly guided by the supervision signals derived from the final objective. Therefore, decoupling the optimization of the encoder and decoder can help mitigate the adverse effects of incorrect pseudo-labels in a semi-supervised setting. This separation allows the encoder to focus on learning robust feature representations while preventing the decoder from being swayed by inaccurate pseudo-labels. Recently, masked image modeling (MIM) has demonstrated robust representation learning capabilities through self-supervised prediction of masked signals, with informed masking proving particularly advantageous (Kakogeorgiou et al., 2022; Wang et al., 2023; Li et al., 2021). Considering the inherent ability of pseudo-labels to identify significant regions worth learning, we ponder whether it is feasible to learn unlabeled data in a pseudo-label-guided MIM fashion.

In this paper, we propose a novel SSL method called PredICt-and-MasK (PICK) for semi-supervised medical image segmentation. PICK distinguishes itself from existing SSL methods in three aspects (see Fig. 1c). First, PICK does not rely on the quality of pseudo-labels or prioritize consistent learning. Instead, it utilizes a primary decoder regulated solely by ground truth to mitigate the impact of incorrect pseudo-labels. This primary decoder identifies target regions to be masked within unlabeled data. Second, PICK seamlessly integrates MIM into SSL through a pseudo-labeled region reconstruction task. This empowers the encoder to deliberately enhance its ability to discern

targets, thereby leveraging an improved understanding of foreground semantics. Third, acknowledging the conflicting objectives between reconstruction and segmentation tasks, PICK introduces an auxiliary decoder that learns from reconstructed images to make semantic predictions. This approach not only enhances the quality of reconstructed images but also reinforces the principle that 'reconstruction serves segmentation' within the model. Under this predict-and-mask strategy, PICK achieves superior performance compared to leading SSL methods across various 2D and 3D medical image segmentation tasks, including the segmentation of lung tumors, pancreas, left atrium, and multiple abdominal organs and domain-generalized tasks. For instance, in lung tumor segmentation, PICK surpasses the best competitor, CauSSL (Miao et al., 2023), by 2.46% on the Dice Coefficient when utilizing 20% labeled data (see Fig. 1d). This demonstrates the efficacy of PICK in leveraging unlabeled data effectively while achieving state-of-the-art segmentation accuracy.

The main contributions of this work are three-fold.

- PICK introduces a novel approach of pseudo-label-guided region masking and reconstruction for mining unlabeled data, distinguishing it from current SSL methods focused on pseudo-labeling or consistency-based co-training.
- Addressing the challenge of conflicting optimization between segmentation and reconstruction, PICK advocates learning from reconstructed images. This enables MIM to seamlessly contribute to SSL while enhancing data diversity through the use of reconstructed images for segmentation.
- Extensive experiments across five public datasets covering diverse segmentation tasks validate PICK's superiority, demonstrating significant improvements over leading SSL methods.

2 Related Work

2.1 SSL Methods

SSL (Chen et al., 2022) involves leveraging both labeled and unlabeled data simultaneously, with the latter typically being more abundant. The core objective of SSL is to effectively transfer learned representations from labeled to unlabeled data. Key strategies employed include pseudo-labeling, consistency-based co-training, uncertainty-based adversarial learning and their hybrids.

Pseudo-labeling aims to identify reliable pseudo-labels for subsequent training iterations. For instance, collaborativemean-teachers have been utilized to improve pseudo-label equality (Tarvainen & Valpola, 2017; Shen et al., 2023). Studies have also explored the trade-off between pseudo-label quality and thresholding based on probability distributions. In our previous work, we introduced a scheme for selecting clean samples based on loss distribution (Zeng et al., 2023). Unlike methods that attempt to filter out unreliable predictions entirely, our approach retains a primary decoder trained solely on labeled data. To harness the potential of unlabeled data, we propose an attentive region masking and reconstruction strategy to enhance the encoder's ability to perceive target features.

Consistency-based co-training seeks to achieve invariant representations by leveraging diverse sub-networks to learn from different perspectives in a co-training framework. For example, differential decoders have been developed to generate consistent predictions (Wu et al., 2022; Zeng et al., 2023; Miao et al., 2023), and discriminative information has been learned through robust constraints under adversarial perturbations (Miyato et al., 2018; Wu et al., 2022; Li et al., 2020). While effective, these methods often require complex designs to maintain inconsistency, thus avoiding the degradation of co-training into vanilla self-training. In contrast, our approach diverges from imposing strict constraints on models to produce complementary predictions. Instead, we propose a novel method where three task-specific but interdependent decoders learn from unlabeled data through a reconstruction proxy task guided by pseudo-labels.

Uncertainty-based adversarial learning aims to improve prediction quality on unlabeled data by identifying reliable target regions during each training step. These methods typically involve generating perturbations (Yang et al., 2023) or adversarial examples to challenge the model, thereby increasing its robustness (Wu & Zhuang, 2022; Hung et al., 2018). For example, Monte-Carlo Dropout (MCDO) (Gal & Ghahramani, 2016) is often employed to assess epistemic uncertainty by performing multiple forward passes, with the resulting uncertainty maps used to incrementally incorporate reliable target regions into the consistency loss (Yu et al., 2019). Some studies (Zeng et al., 2023; Miyato et al., 2018) inject adversarial perturbations at both the image and feature levels to identify and exploit discriminative information. However, these methods often require specialized designs for uncertainty estimation and perturbation simulation, which can lead to issues such as increased model complexity, potential instability, and higher computational costs.

2.2 SSL for Medical Image Analysis

In medical imaging, SSL methods have gained significant attention for their ability to leverage limited labeled data alongside larger pools of unlabeled data, thus enhancing performance while reducing the need for extensive man-

ual labeling. For instance, Wu and Zhuang (2022) proposed a risk minimization framework that establishes connections between labeled and unlabeled data using an unbiased estimator. Wang and Li (2023) introduced a re-weighting strategy that takes into account both class distribution and segmentation difficulty to enhance the accuracy of multi-organ segmentation. Zeng et al. (2024) advocated differential decoder feature learning to improve the effectiveness of unlabeled data utilization, with a focus on handling ambiguous organ and tumor regions. Additionally, Wang et al. (2023) focused on improving pseudo-label quality and addressing cognitive biases by utilizing inconsistent predictions between sub-networks to guide corrections.

In comparison to these existing methods, our proposed PICK framework makes novel contributions to the SSL community. Notably, PICK embeds unlabeled target information into the encoder through a pseudo-label guided MIM task. In contrast, existing SSL methods primarily use pseudo-label supervision to enhance decoder performance, which can inadvertently introduce incorrect predictions. Our design, however, ensures that the task-relevant decoder remains unaffected by these inaccuracies.

2.3 MIM for Visual Representation

MIM involves predicting masked signals based on visible, unmasked signals (He et al., 2022; Xie et al., 2022). It serves as a potent pretext task in self-supervised learning, aiming to enhance representation learning capabilities. Recent advancements in MIM emphasize the benefits of informed masking over random masking. For instance, Li et al. (2021) dynamically masked foreground patches using attention maps, Kakogeorgiou et al. (2022) distilled a student model by predicting masked objects highlighted by a teacher model, and Wang et al. (2023) introduced a challenging masking strategy via hard patch mining. These studies highlight the efficacy of attentive masking strategies in extracting rich feature representations from visual data.

In the context of SSL, pseudo-labels generated for unlabeled data inherently indicate regions of interest for learning (Caron et al., 2021). This observation inspires the integration of MIM into SSL frameworks. However, integrating MIM poses challenges due to conflicting optimization objectives between segmentation and reconstruction tasks. To resolve this challenge, we propose the inclusion of an auxiliary decoder. This auxiliary decoder is tailored to harmonize the objectives of segmentation and reconstruction within our SSL method, ensuring effective utilization of unlabeled data while maintaining segmentation performance.

3 Method

3.1 Preliminaries

In the realm of SSL, the training dataset $\mathcal{D} = \{\mathcal{D}_l, \mathcal{D}_u\}$ typically contains a labeled subset $\mathcal{D}_l = \left\{(x_i^l, y_i^l)\right\}_{i=1}^{N_l}$ and an unlabeled subset $\mathcal{D}_u = \left\{(x_i^u)\right\}_{i=1}^{N_u}$, where x_i^l and y_i^l denote the i-th labeled image and its corresponding annotation, respectively, while x_i^u represents the i-th unlabeled image. The subsets have cardinalities N_l and N_u, where with $N_l \ll N_u$. The objective of SSL is to effectively utilize both $\mathcal{D}_l$ and $\mathcal{D}_u$ to train a model that performs well on unseen test sets. The distinction among SSL approaches primarily resides in their unique exploitation $\mathcal{D}_u$.

3.2 Overview of PICK Framework

Our proposed PICK framework, depicted in Fig. 2, integrates an encoder $f_E(\cdot;\theta)$ with three distinct decoders: a primary decoder $f_D(\cdot;\theta_{pri})$, a MIM decoder $f_D(\cdot;\theta_{mim})$, and an auxiliary decoder $f_D(\cdot;\theta_{aux})$. Each decoder shares the same architecture but serves different objectives. Specifically, The primary decoder, trained exclusively on labeled data, is expected to generate pseudo-labels for the unlabeled data. Its role is to identify target regions within the unlabeled images. The MIM decoder is dedicated to reconstructing the target regions highlighted by the pseudo-labels generated by the primary decoder. This decoder enhances the encoder's ability to understand target semantics through reconstruction. The auxiliary decoder is crafted to learn from the reconstructed images, by aligning the segmentation results with pseudo-labels. It helps in harmonizing the conflicting objectives of segmentation and reconstruction tasks. We now delve into the details of training PICK.

3.3 Warm-up on Labeled Data

In SSL, a common practice is to initially train the model on labeled data to establish a solid foundation for handling unlabeled data effectively (Zeng et al., 2023). This approach, referred to as warm-up, aims to equip the model with robust generalization capabilities.

In the context of PICK, the warm-up phase begins with the primary decoder $f_D(\cdot;\theta_{pri})$ being trained solely on the labeled subset $\mathcal{D}_l = \left\{(x_i^l, y_i^l)\right\}_{i=1}^{N_l}$. The primary decoder $f_D(\cdot;\theta_{pri})$ predicts pseudo-labels $\hat{y}_{pri}^l$ for each x_i^l, indicating regions of interest (e.g., target tumors or organs) based on the model's current understanding derived from the labeled data. This process can be formulated as:

$$\hat{y}_{pri}^l = \arg\max(f_D(feat^l;\theta_{pri})), \tag{1}$$

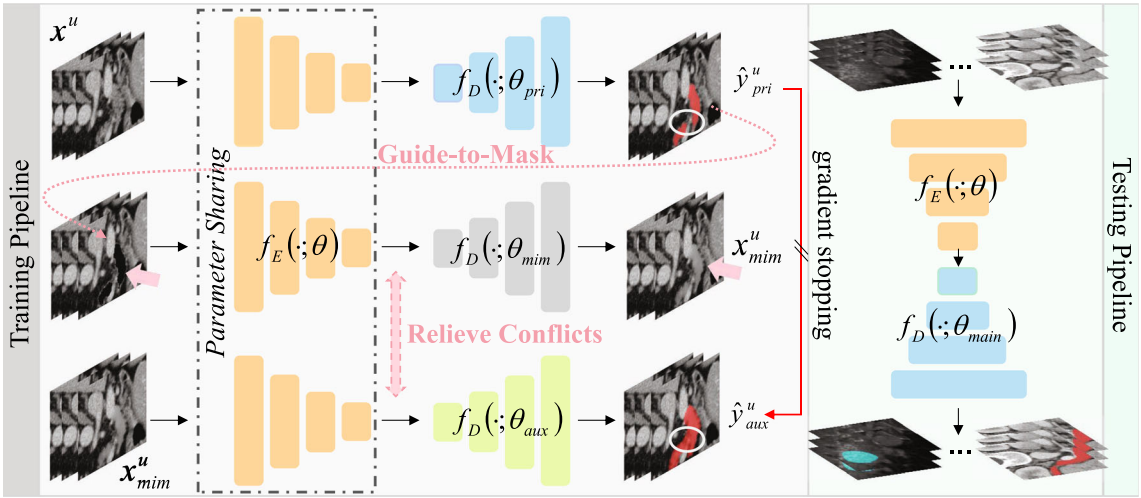

Fig. 2 Training and testing pipelines of our proposed PICK. PICK utilizes a shared encoder and three task-specific decoders. Given an unlabeled image x^u, the primary decoder $f_D(\cdot;\theta_{pri})$ generates a pseudo-label $\hat{y}^u_{pri}$ indicating the region-of-interest, guiding the masking of x^u. The masked image modeling decoder $f_D(\cdot;\theta_{mim})$ reconstructs masked regions to enhance target semantics through a reconstruction task. To mitigate the conflicts between segmentation and reconstruction, an auxiliary decoder $f_D(\cdot;\theta_{aux})$ predicts $\hat{y}^u_{aux}$ on the reconstructed image x^u_{mim}, supervised by $\hat{y}^u_{pri}$. Note that $f_D(\cdot;\theta_{pri})$ is solely supervised by labeled data and used for inference only. In this illustration, the masked regions are indicated by the pink arrows, and the inconsistent regions between $\hat{y}^u_{pri}$ and $\hat{y}^u_{aux}$ are highlighted by white ellipses

where $feat^l = f_E(x^l;\theta)$ represents the feature embedding of the labeled image x^l. These pseudo-labels indicate the regions of interest (e.g., target tumors or organs) based on the model's current understanding from the labeled data.

Subsequently, x^l_i is masked using the predicted pseudo-labels $\hat{y}^l_{pri}$. This process involves overlaying the pseudo-labels onto x^l_i to highlight the regions identified by $f_D(\cdot;\theta_{pri})$ as important for learning. The masked image x^l_{mask} is then employed to train the MIM decoder $f_D(\cdot;\theta_{mim})$. The above process can be written as:

$$x^l_{mask} = x^l \odot (1 - \hat{y}^l_{pri}), \tag{2}$$

$$x^l_{mim} = f_D(feat^l_{mask};\theta_{mim}), \tag{3}$$

where $feat^l_{mask} = f_E(x^l_{mask};\theta)$ is the feature embedding of x^l_{mask}. The role of $f_D(\cdot;\theta_{mim})$ is to reconstruct the masked regions based on the information encoded in x^l_i and $\hat{y}^l_{pri}$. This process encourages the encoder $f_E(\cdot;\theta)$ to learn more discriminative features relevant to the target regions indicated by the pseudo-labels.

To address the inherent conflict between segmentation and reconstruction tasks, PICK introduces the auxiliary decoder $f_D(\cdot;\theta_{aux})$. This decoder learns from the reconstructed image x^l_{mim} produced by $f_D(\cdot;\theta_{mim})$, aligning the segmentation prediction with the pseudo-label. This process can be expressed in a formula as:

$$\hat{y}^l_{aux} = \arg\max(f_D(feat^l_{mim};\theta_{aux})), \tag{4}$$

where $feat^l_{mim} = f_E(x^l_{mim};\theta)$ represents the feature embedding of the reconstructed labeled image x^l_{mim}. By supervising $f_D(\cdot;\theta_{aux})$ on those reconstructed images, PICK ensures that the segmentation task benefits from the enhanced feature representations derived from the MIM reconstruction. Note that, during warm-up, we utilize the predicted pseudo-labels $\hat{y}^l_{pri}$, instead of the ground truth annotations y^l_i, to mask labeled images for two reasons: (1) predictions tend to converge towards the ground truth during supervised learning, and (2) predictions often exhibit more variability and incorporate the model's uncertainty and insights, which can be beneficial for learning robust representations.

3.4 Predict-and-Mask on Unlabeled Data

Target Region Revelation with $f_D(\cdot;\theta_{pri})$. Unlike conventional pseudo-labeling approaches that risk perpetuating unreliable predictions, our primary decoder $f_D(\cdot;\theta_{pri})$ is exclusively trained on labeled data. This strategy ensures that PICK maintains a clean and highly relevant decoder specific to the segmentation task. When processing unlabeled data $\mathcal{D}_u = \{(x^u_i)\}_{i=1}^{N_u}$, this decoder identifies the attentive regions of targets, denoted as $\hat{y}^u_{pri}$, computed as

$$\hat{y}^u_{pri} = \arg\max(f_D(feat^u;\theta_{pri})), \tag{5}$$

where $feat^u = f_E(x^u;\theta)$ represents the feature embedding of the unlabeled image x^u.

Attentive Region Reconstruction with $f_D(\cdot;\theta_{mim})$. Upon identifying $\hat{y}^u_{pri}$, which may contain errors, traditional

pseudo-labeling and co-training methods might propagate these errors into the decoder through cross-pseudo supervision. To extract valuable information in $\hat{y}^u_{pri}$ while minimizing error propagation, we adopt a strategy that masks and reconstructs the target regions indicated by $\hat{y}^u_{pri}$. This process is formalized as follows

$$x^u_{mask} = x^u \odot (1 - \hat{y}^u_{pri}), \tag{6}$$
$$x^u_{mim} = f_D(feat^u_{mask}; \theta_{mim}), \tag{7}$$

where x^u_{mask} and x^u_{mim} denote the masked and reconstructed images, respectively, and $feat^u_{mask} = f_E(x^u_{mask}; \theta)$ is the feature embedding of x^u_{mask}. The element-wise multiplication is denoted by $\odot$. The MIM decoder is guided by a masked image reconstruction loss

$$\mathcal{L}^u_{mim} = \frac{1}{\Omega(\hat{y}^u_{pri})} \|\hat{y}^u_{pri}(x^u_{mim} - x^u)\|_1, \tag{8}$$

where $\Omega(\cdot)$ denotes the number of elements. Despite bolstering the information within $\hat{y}^u_{pri}$, a disparity between segmentation and reconstruction tasks necessitates the use of an auxiliary decoder $f_D(\cdot; \theta_{aux})$.

This MIM proxy task allows PICK to leverage unlabeled data to enrich the encoder's knowledge (validated in Sect. 5.3 with visualizations), while preserving a reliable decoder $f_D(\cdot; \theta_{pri})$ constrained by accurate supervision alone.

Integration of Segmentation and Reconstruction with $f_D(\cdot; \theta_{aux})$. The divergent goals of segmentation and reconstruction pose challenges for $f_E(\cdot; \theta)$ in generating optimal features. To address this, we propose learning from x^u_{mim} again, where segmentation not only reflects reconstruction quality but also emphasizes the ultimate utility of reconstruction for segmentation. Moreover, reconstructed images can augment source data diversity, enhancing segmentation robustness. Here, $f_D(\cdot; \theta_{aux})$ is supervised by $\hat{y}^u_{pri}$ and is inactive during inference. The auxiliary loss $\mathcal{L}^u_{aux}$ is defined as

$$\mathcal{L}^u_{aux} = \mathcal{L}_{seg}(f_D(feat^u_{mim}; \theta_{aux}), \hat{y}^u_{pri}), \tag{9}$$

where $\mathcal{L}_{seg}(\cdot) = \mathcal{L}_{Dice}(\cdot) + \mathcal{L}_{CE}(\cdot)$, $feat^u_{mim} = f_E(x^u_{mim}; \theta)$, $\mathcal{L}_{Dice}(\cdot)$ is the Dice loss, and $\mathcal{L}_{CE}(\cdot)$ is the cross entropy loss. Through coordinated use of these task-specific decoders, PICK achieves nuanced understanding and effective utilization of unlabeled data.

3.5 Training and Testing

3.5.1 Training

Section 3.4 provides the details of learning from unlabeled data. The unsupervised loss $\mathcal{L}_{unsup}$ calculated on unlabeled data is defined as

$$\mathcal{L}_{unsup} = \lambda \mathcal{L}^u_{mim} + \mathcal{L}^u_{aux}, \tag{10}$$

where λ is a trade-off coefficient. For labeled data, the learning procedure mirrors that of unlabeled data, employing a predict-and-mask approach. The only difference is that the ground truth is employed as the supervision signal. The supervised loss $\mathcal{L}_{sup}$ calculated on labeled data is formulated as

$$\begin{aligned} \mathcal{L}_{sup} &= \lambda \mathcal{L}^l_{mim}(f_D(feat^l_{mask}; \theta_{mim}), x^l) \\ &\quad + \mathcal{L}^l_{aux}(f_D(feat^l_{mim}; \theta_{aux}), y^l) \\ &\quad + \mathcal{L}_{seg}(f_D(feat^l; \theta_{pri}), y^l), \end{aligned} \tag{11}$$

where $feat^l = f_E(x^l; \theta)$, $feat^l_{mask} = f_E(x^l_{mask}; \theta)$ and $feat^l_{mim} = f_E(x^l_{mim}; \theta)$, which are feature embedding of x^l, x^l_{mask} and x^l_{mim}, respectively. The overall loss $\mathcal{L}$ is written as:

$$\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{unsup}. \tag{12}$$

3.5.2 Testing

Given an unseen test image, PICK firstly generates features using the encoder $f_E(\cdot; \theta)$, and then uses the primary decoder $f_D(\cdot; \theta_{pri})$ to perform segmentation.

4 Experiments and Results

4.1 Datasets

To evaluate the efficacy of our PICK framework, we conducted extensive experiments on five public medical image segmentation datasets. These datasets cover a range of segmentation tasks, including lung tumor and pancreas segmentation from CT scans, left atrium segmentation from MRI images, multi-organ segmentation from CT scans, and 2D cardiac segmentation from multi-vendor data. Here are the details of these datasets.

MSD-Lung Tumor: This dataset (Antonelli et al., 2022) contains 63 CT scans. We employed a 4-fold cross-validation scheme for assessment, using randomly cropped patches of size 96×96×96 for evaluation.

NIH-Pancreas: This dataset (Roth et al., 2015) consists of 82 contrast-enhanced CT scans. We followed the evaluation protocol in Wang et al. (2023) and conducted a 4-fold cross-validation, with input patches of size 96×96×96.

Left Atrium: Containing 100 gadolinium-enhanced MRI images, this dataset (Xiong et al., 2021) was subjected to a 5-

Table 1 4-fold cross-validation results (mean±std) on the MSD-Lung Tumor and NIH-Pancreas dataset, using 20% labeled data

Method	MSD-Lung Tumor (20%/10 labeled data)				NIH-Pancreas (20%/12 labeled data)			
	Dice (%) ↑	Jaccard (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓	Dice (%) ↑	Jaccard (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓
VNet (Baseline)	36.36 ± 1.450	25.78 ± 0.972	19.06 ± 1.642	32.73 ± 1.347	64.18 ± 0.073	49.26 ± 0.077	4.69 ± 0.935	17.74 ± 3.572
UA-MT (Yu et al., 2019)	44.72 ± 0.631	31.20 ± 0.969	11.36 ± 0.827	24.36 ± 1.733	74.01 ± 0.029	60.00 ± 3.031	5.19 ± 1.267	17.00 ± 3.031
SASSNet (Li et al., 2020)	45.51 ± 0.780	31.76 ± 0.883	9.76 ± 0.762	20.08 ± 1.631	73.57 ± 0.017	59.71 ± 0.020	3.53 ± 1.416	13.87 ± 1.079
DTC (Luo et al., 2021)	49.01 ± 1.060	35.50 ± 0.782	7.78 ± 1.376	21.12 ± 2.012	73.23 ± 0.024	59.18 ± 0.027	3.81 ± 0.953	13.20 ± 2.241
MC-Net (Wu et al., 2022)	50.20 ± 0.782	36.47 ± 0.529	5.02 ± 0.898	18.33 ± 1.376	73.73 ± 0.019	59.19 ± 0.021	3.92 ± 1.055	13.65 ± 3.902
CAML (Gao et al., 2023)	53.78 ± 0.668	37.92 ± 0.358	4.10 ± 1.262	13.25 ± 0.780	76.57 ± 0.525	63.10 ± 0.442	*3.03 ± 0.822*	12.57 ± 1.052
UCMT (Shen et al., 2023)	54.37 ± 0.502	38.77 ± 0.380	3.76 ± 0.106	12.06 ± 0.202	76.37 ± 0.753	62.77 ± 0.538	3.76 ± 0.526	12.06 ± 0.702
MCF (Wang et al., 2023)	52.62 ± 0.724	36.62 ± 0.837	5.72 ± 0.336	14.43 ± 0.572	75.00 ± 0.026	61.27 ± 0.030	3.27 ± 0.919	11.59 ± 1.611
CauSSL (Miao et al., 2023)	*55.19 ± 0.754*	*39.97 ± 0.721*	*3.52 ± 0.473*	*11.97 ± 0.762*	*77.86 ± 0.639*	*64.57 ± 0.562*	3.47 ± 0.362	*9.73 ± 0.747*
PICK (Ours)	**57.65 ± 0.268**	**42.40 ± 0.470**	**2.76 ± 0.212**	**10.85 ± 0.436**	**80.14 ± 0.214**	**67.65 ± 0.309**	**1.55 ± 0.128**	**6.12 ± 0.422**

The best and second best results are highlighted in bold and italic, respectively

Table 2 5-fold cross-validation results (mean ± std) on the Left Atrium dataset, using 10% and 20% labeled data

Method	Left Atrium (10%/8 labeled data)				Left Atrium (20%/16 labeled data)			
	Dice (%) ↑	Jaccard (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓	Dice (%) ↑	Jaccard (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓
VNet (Baseline)	78.62 ± 0.992	68.45 ± 0.763	4.68 ± 0.808	18.45 ± 1.630	83.34 ± 0.023	72.49 ± 0.029	3.87 ± 0.337	14.77 ± 1.169
UA-MT (Yu et al., 2019)	82.78 ± 0.872	72.37 ± 0.891	4.31 ± 1.380	17.45 ± 1.882	85.98 ± 0.014	76.65 ± 0.017	2.68 ± 0.776	9.86 ± 2.707
SASSNet (Li et al., 2020)	84.02 ± 0.720	73.46 ± 0.606	3.78 ± 0.741	13.28 ± 0.897	86.21 ± 0.023	77.15 ± 0.024	2.68 ± 0.416	9.80 ± 1.842
DTC (Luo et al., 2021)	84.38 ± 0.881	74.27 ± 0.734	4.02 ± 1.078	14.97 ± 1.768	86.36 ± 0.023	77.25 ± 0.020	2.40 ± 0.223	9.02 ± 1.015
MC-Net (Wu et al., 2022)	85.48 ± 0.782	75.60 ± 0.609	3.83 ± 0.766	13.39 ± 1.022	87.65 ± 0.011	78.63 ± 0.013	3.01 ± 0.700	9.70 ± 2.361
CAML (Gao et al., 2023)	*87.76 ± 0.582*	*78.21 ± 0.873*	*2.69 ± 0.388*	*10.33 ± 0.603*	89.27 ± 0.412	80.98 ± 0.329	2.36 ± 0.270	7.38 ± 0.626
UCMT (Shen et al., 2023)	87.23 ± 0.557	78.03 ± 0.742	3.01 ± 0.336	10.98 ± 0.570	88.38 ± 0.429	80.20 ± 0.336	2.02 ± 0.433	9.28 ± 0.981
MCF (Wang et al., 2023)	86.65 ± 0.623	77.41 ± 0.830	2.78 ± 0.400	12.95 ± 0.733	88.71 ± 0.018	80.41 ± 0.022	*1.90 ± 0.187*	*6.32 ± 0.800*
CauSSL (Miao et al., 2023)	87.03 ± 0.535	77.92 ± 0.778	3.12 ± 0.216	11.27 ± 0.622	*89.39 ± 0.334*	*81.07 ± 0.427*	2.38 ± 0.366	8.65 ± 1.021
PICK (Ours)	**88.58 ± 0.574**	**79.69 ± 0.462**	**2.32 ± 0.167**	**9.74 ± 0.828**	**90.85 ± 0.163**	**83.37 ± 0.224**	**1.74 ± 0.128**	**5.70 ± 0.282**

The best and second best results are highlighted in bold and italic, respectively

fold cross-validation to ensure rigorous comparison. Training patches were randomly cropped to a size of 80×112×112.

AMOS: This is an abdominal multi-organ dataset (Ji et al., 2022) with 360 scans covering 15 organs, including the spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, and prostate/uterus. Following the data split in Wang and Li (2023), we partitioned the dataset into training (216 scans), validation (24 scans), and test (120 scans) sets. During training, the input size was standardized to 64×96×96.

M&Ms: This dataset Campello et al. (2021) features 320 subjects scanned by various vendors (GE, Canon, Philips, Siemens). Following (Wang & Li, 2023), we divided the data into four domains (A, B, C, and D) based on vendors, with subject counts of 95, 125, 50, and 50, respectively. We adopted the leave-one-domain-out approach, where one domain serves as the testing set while the model is trained on other three domains. For example, the results for Domain A are derived from a model trained on Domains B, C, and D. The images were resized to 288×288, making M&Ms suitable as a 2D semi-supervised domain-generalized dataset for our experiments.

4.2 Implementation Details

Following established practices (Luo et al., 2021; Wang et al., 2023; Miao et al., 2023), we implemented PICK using V-Net (Milletari et al., 2016) for 3D tasks and U-Net (Ronneberger et al., 2015) for 2D tasks as our baseline models. The encoder architecture of PICK consists of five down-sampling stages, incorporating a total of 16 convolutional layers that produce feature maps with channel dimensions progressing through 16, 32, 64, 128, and 256. Each of the three decoders shares an identical architecture, comprising five up-sampling stages with 14 convolutional layers. For optimization, we employed the SGD optimizer with an initial learning rate of 0.01, momentum of 0.9, and weight decay fixed at 1e-4. Our training regimen utilized a batch size of 8, consisting of 4 labeled and 4 unlabeled data samples. Data augmentation techniques included random rotation, flipping, and (Yun et al., 2019). Performance evaluation of PICK utilized standard metrics such as Dice coefficient (%), Jaccard index (%), Average Surface Distance (ASD, in voxels), and 95% Hausdorff Distance (95HD, in voxels). All experiments were conducted using Paszke et al. (2019) on a workstation equipped with four NVIDIA GeForce RTX 3080 Ti GPUs.

Table 3 Performance comparison on the AMOS dataset using 5% annotations. The 15 classes are spleen (Sp), right kidney (RK), left kidney (LK), gallbladder (Ga), esophagus (Es), liver(Li), stomach(St), aorta (Ao), inferior vena cava (IVC), pancreas (Pa), right adrenal gland (RAG), left adrenal gland (LAG), duodenum (Du), bladder (Bl), and prostate/uterus (P/U)

Methods	Avg.		Average Dice of Each Class													
	Dice	ASD	Sp	RK	LK	Ga	Es	Li	St	Ao	IVC	Pa	RAG	LAG	Du	Bl	P/U
UA-MT (Yu et al., 2019)	42.16	15.48	59.8	64.9	64.0	35.3	34.1	77.7	37.8	61.0	46.0	33.3	26.9	12.3	18.1	29.7	31.6
CPS (Chen et al., 2021)	41.08	20.37	56.1	60.3	59.4	33.3	25.4	73.8	32.4	65.7	52.1	31.1	25.5	6.2	18.4	40.7	35.8
DST (Chen et al., 2022)	41.44	21.12	58.9	63.3	63.8	37.7	29.6	74.6	36.1	66.1	49.9	32.8	13.5	5.5	17.6	39.1	33.1
DePL (Wang et al., 2022)	41.97	20.42	55.7	62.4	57.7	36.6	31.3	68.4	33.9	65.6	51.9	30.2	23.3	10.2	20.9	43.9	37.7
Adsh (Guo et al., 2022)	40.33	24.53	56.0	63.6	57.3	34.7	25.7	73.9	30.7	65.7	51.9	27.1	20.2	0.0	18.6	43.5	35.9
CReST (Wei et al., 2021)	46.55	14.62	66.5	64.2	65.4	36.0	32.2	*77.8*	43.6	68.5	52.9	40.3	24.7	19.5	26.5	43.9	36.4
SimiS (Chen et al., 2022)	47.27	11.51	**77.4**	**72.5**	68.7	32.1	14.7	86.6	*46.3*	74.6	54.2	41.6	24.4	17.9	21.9	**47.9**	28.1
CLD (Lin et al., 2022)	46.10	15.86	67.2	68.5	71.4	41.0	21.0	76.1	42.4	69.8	52.1	37.9	24.7	23.4	22.7	38.1	35.2
DHC (Wang & Li, 2023)	49.53	13.89	68.1	69.6	71.7	*42.3*	37.0	76.8	43.8	70.8	*57.4*	43.2	*27.0*	28.7	**29.1**	41.4	36.7
MagicNet (Chen et al., 2023)	*50.41*	*11.07*	72.9	70.8	*73.0*	41.9	*38.2*	74.3	44.9	*75.0*	56.8	*44.8*	26.6	*29.2*	27.6	43.0	*37.1*
PICK (Ours)	**51.96**	**9.78**	*74.3*	*71.9*	**74.2**	**44.6**	**39.5**	**77.8**	**47.2**	**75.2**	**58.2**	**45.7**	**28.8**	**30.1**	*28.3*	*47.2*	**37.8**

The best and second best results are highlighted in bold and italic, respectively. Averaged Dice and ASD scores over 15 classes are reported

Table 4 Performance comparison on the **AMOS** dataset using 10%

Methods	Avg.		Average dice of each class													
	Dice	ASD	Sp	RK	LK	Ga	Es	Li	St	Ao	IVC	Pa	RAG	LAG	Du	Bl	P/U
UA-MT (Yu et al., 2019)	40.60	38.45	61.0	75.4	58.8	0.1	0.0	84.4	45.2	72.8	61.6	36.2	0.0	0.0	30.7	46.5	36.3
CPS (Chen et al., 2021)	54.51	29.69	80.7	79.8	74.3	35.2	44.4	90.5	51.5	76.1	65.6	48.6	31.6	21.8	33.6	47.0	37.3
DST (Chen et al., 2022)	52.24	17.66	81.7	80.2	78.6	39.5	41.0	89.8	52.8	78.5	65.9	51.1	4.3	0.1	34.2	48.8	37.2
DePL (Wang et al., 2022)	56.76	6.70	81.9	80.6	79.5	41.0	42.6	89.3	57.6	79.1	66.0	53.2	34.6	21.8	34.9	48.4	40.9
Adsh (Guo et al., 2022)	54.92	8.07	81.6	78.5	76.6	40.1	43.4	90.1	53.0	76.7	64.4	48.3	25.9	24.2	34.7	48.7	37.7
CReST (Wei et al., 2021)	60.74	4.65	85.3	84.5	84.0	43.2	50.8	89.9	58.7	84.7	73.0	54.2	41.8	31.6	41.0	52.8	35.8
SimiS (Chen et al., 2022)	57.48	4.46	83.1	80.9	80.0	39.6	45.9	90.0	57.1	78.0	66.3	54.1	35.8	26.9	39.9	49.3	35.4
CLD (Lin et al., 2022)	61.55	4.21	86.0	85.3	84.8	44.5	51.9	90.8	59.7	83.7	73.1	55.7	40.2	37.2	41.4	53.0	36.1
DHC (Wang & Li, 2023)	64.16	3.51	*87.4*	86.6	**87.1**	45.8	57.0	89.8	**64.7**	86.0	75.0	62.5	39.8	36.8	44.0	*56.5*	43.6
MagicNet (Chen et al., 2023)	*64.89*	*3.18*	86.7	*87.2*	*86.3*	*46.5*	*57.7*	*91.2*	63.3	*87.1*	*76.2*	*62.9*	*41.2*	*40.7*	*45.1*	56.4	*44.9*
PICK (Ours)	**66.37**	**2.65**	**88.8**	**89.4**	86.2	**47.3**	**59.6**	**91.7**	*64.0*	**88.3**	**77.9**	**64.1**	**43.5**	**43.0**	**47.2**	**58.8**	**45.7**

annotationsThe best and second best results are highlighted in bold and italic, respectively. Averaged Dice and ASD scores over 15 classes are reported

4.3 Comparison with State of the Art

4.3.1 Results on MSD-Lung Tumor Dataset

We compared the proposed PICK framework with eight leading SSL methods on the MSD-Lung Tumor dataset, namely UA-MT (Yu et al., 2019), SASSNet (Li et al., 2020), DTC (Luo et al., 2021), MC-Net (Wu et al., 2022), CAML (Gao et al., 2023), UCMT (Shen et al., 2023), MCF (Wang et al., 2023), and CauSSL (Miao et al., 2023). To simulate the SSL setting, we randomly selected 20% of training samples (10 samples) as labeled data and left the others as unlabeled data. Table 1 (left) summarizes the mean and standard deviation of performance metrics. It shows that PICK outperforms UCMT, the leading pseudo-labeling method that utilizes two teachers to collaboratively supervise a student that thereby mitigates the impact of incorrect pseudo-labels, by 3.28% in Dice score and 1.21 voxels in 95HD. This improvement highlights the advantage of maintaining a clean decoder, which minimizes the influence of erroneous pseudo-labels. Additionally, compared to MCF, which employs a contrastive discrepancy review design to reduce the impact of biased predictions, PICK presents a 5.03% improvement in Dice score. This emphasizes the effectiveness of embedding unlabeled target information into the encoder while preserving an unaffected primary decoder. Notably, PICK also outperforms the best-performing co-training method, CauSSL, by 2.46% in Dice score and 1.12 voxels in 95HD. This highlights the effectiveness of our predict-and-mask strategy, which simplifies co-training by mitigating the impact of incorrect pseudo-labels. Visual inspection in Fig. 3 (first row) reveals the challenges faced by other SSL methods in generating complete segmentations, particularly in tumor segmentation tasks. Despite these challenges, PICK demonstrates promising segmentation results compared to its competitors.

4.3.2 Results on NIH-Pancreas Dataset

Table 1 (right) presents the performance comparison of PICK against the aforementioned eight competing methods on the NIH-Pancreas dataset, where 20% of training samples (12 samples) are labeled. It reveals that PICK achieves the highest scores across all metrics, surpassing the second-best method, CauSSL, by 2.28% in Dice score and 3.61 voxels in 95HD. This demonstrates the superiority of our approach over prevalent consistency-based co-training methods. Additionally, PICK outperforms the bias correction method MCF by 5.14% in Dice score, highlighting the efficacy of maintaining a primary decoder supervised solely by ground truth data to mitigate the impact of incorrect pseudo-labels. Similarly, PICK surpasses the uncertainty-guided teacher collaborative framework UCMT by 3.77% in Dice score. These improvements underscore the efficacy of

Table 5 Dice performance on the M&Ms dataset, using 2% and 5% labeled data

Method	2% label percentage					5% label percentage				
	Domain A	Domain B	Domain C	Domain D	Average	Domain A	Domain B	Domain C	Domain D	Average
U-Net (Baseline)	54.83	54.97	66.83	63.56	60.05	61.09	70.36	70.06	71.03	68.14
LDDG	59.47	56.16	68.21	68.56	63.16	66.22	69.49	73.40	75.66	71.29
SAML (Liu et al., 2020)	56.31	56.32	75.70	69.94	64.57	67.11	76.35	77.43	78.64	74.88
DGNet (Liu et al., 2021)	66.01	72.72	77.54	75.14	72.85	72.40	80.30	82.51	83.77	79.75
vMFNet (Liu et al., 2022)	73.13	77.01	*81.57*	82.02	78.43	77.06	82.29	*84.01*	*85.13*	82.12
BCP (Bai et al., 2023)	71.57	76.20	76.87	77.94	75.65	73.66	79.04	77.01	78.49	77.05
A&D (Wang & Li, 2023)	*79.62*	**82.26**	80.03	*83.31*	*81.31*	*81.71*	**85.44**	82.18	83.90	*83.31*
PICK (Ours)	**80.70**	*81.92*	**83.58**	**83.87**	**82.52**	**82.63**	*84.72*	**84.73**	**85.20**	**84.32**

The best and second best results are highlighted in bold and italic, respectivelyResults on Domain A are produced by the model trained on Domains B, C, and D

Fig. 3 Visualization results on 3D datasets. Competitive methods including MC-Net (Wu et al., 2022), CauSSL (Miao et al., 2023), CPS, and MagicNet (Chen et al., 2023) are listed for comparison

our predict-and-mask design, which leverages unlabeled data through a pseudo-label guided MIM pretext task. Moreover, the second row of Fig. 3 illustrates pancreas segmentation results, where PICK produces the most accurate segmentation compared to other methods.

4.3.3 Results on Left Atrium Dataset

We conducted experiments comparing PICK with aforementioned eight competing methods on the Left Atrium dataset using 10% (8 cases) and 20% (16 cases) of training samples as labeled data, respectively, and reported the results in Table 2. It shows that, compared to the attention-based CAML, PICK demonstrates a 1.48% improvement in Jaccard index with 10% labeled data and a reduction of 1.68 voxels in 95HD with 20% labeled data. This improvement can be attributed to PICK's focus on masking and reconstructing attentive regions, which emphasizes the most informative regions without requiring complex attention mechanisms. Compared to UA-MT, which employs uncertainty estimation, PICK achieves a significant 4.87% increase in Dice score, further confirming its robustness. Moreover, compared to MCF and UCMT, which each employ unique designs to prevent erroneous pseudo-labels from directly influencing the model, our PICK reduces 95HD by 3.21 and 1.51 voxels, respectively, when using 10% labeled data. This phenomenon showcases the superior effectiveness of PICK in mitigating the impact of erroneous pseudo-labels, highlighting its robustness in improving segmentation accuracy with limited labels. Visualization in the third row of Fig. 3 illustrates the segmentation results for the left atrium. It is evident that PICK produces the most comprehensive segmentation.

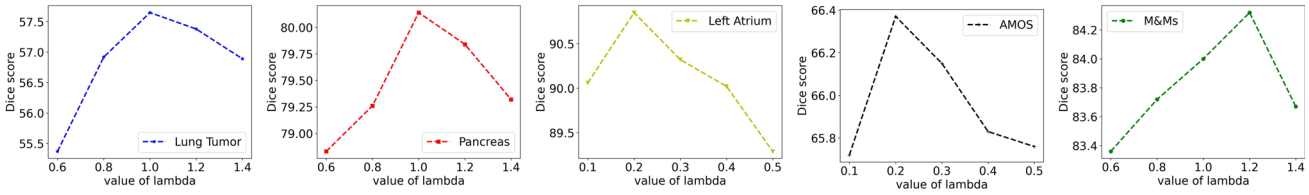

Fig. 4 Analysis of hyper-parameter λ on the MSD-Lung Tumor, NIH-Pancreas, Left Atrium, AMOS, and M&Ms datasets, with labeled data percentages of 20%, 20%, 20%, 10%, and 5%

Table 6 Ablation study of each component in PICK on the MSD-Lung Tumor dataset, NIH-Pancreas dataset, and Left Atrium dataset

$f_D^{\theta_{pri}}$	$f_D^{\theta_{mim}}$	$f_D^{\theta_{aux}}$	MSD-lung tumor dataset				NIH-pancreas dataset				Left atrium dataset			
			Dice	Jaccard	ASD	95HD	Dice	Jaccard	ASD	95HD	Dice	Jaccard	ASD	95HD
√			36.36	25.78	19.06	32.73	64.18	49.26	4.69	17.74	83.34	72.49	3.87	14.77
√	√		53.56	38.28	2.66	9.52	77.15	63.58	1.95	8.45	88.39	80.07	2.83	7.65
√		√	47.36	32.63	1.84	9.98	73.86	60.61	2.71	10.96	86.68	77.65	3.31	8.90
√	√	√	57.65	42.40	2.76	10.85	80.14	67.65	1.55	6.12	90.85	83.37	1.74	5.70

4.3.4 Results on AMOS Dataset

Next, we applied PICK to multi-organ segmentation and compared it against ten leading methods, including UA-MT (Yu et al., 2019), CPS (Chen et al., 2021), DST (Chen et al., 2022), DePL (Wang et al., 2022), Adsh (Guo et al., 2022), CReST (Wei et al., 2021), SimiS (Chen et al., 2022), CLD (Lin et al., 2022), DHC (Wang & Li, 2023), and MagicNet (Chen et al., 2023), on the AMOS dataset. Tables 3 and 4 present the performance of these methods using 5% and 10% of training samples as labeled data, respectively. It shows that, PICK consistently outperforms the relation-based augmentation method, MagicNet, achieving improvements of 1.55% and 1.48% in average Dice score over 15 segmentation targets with 5% and 10% annotations, respectively. These improvements underscore PICK's adaptability without relying on augmentation techniques based on organ relations. Moreover, PICK shows a 2.43% increase in average Dice score compared to the dynamic class reweighting method DHC with 5% labeled data, underscoring the effectiveness of our predict-and-mask strategy. Furthermore, PICK demonstrates a significant 10.52% improvement in Dice score over the debiased self-training method DST. Despite DST employing two independent heads with min-max optimization to mitigate biases, the adversarial loss may cause potential shifts in model learning on labeled data. In contrast, PICK simplifies the handling of biases from incorrect predictions through a pseudo-label guided MIM task, which embeds unlabeled target information into the encoder while keeping the task-specific decoder unaffected by erroneous pseudo-labels. Additionally, as observed from the last row of Fig. 3, PICK effectively identifies small regions overlooked by other methods. This enhancement is attributed to the improved perceptual ability of the encoder, facilitated by pseudo-label-guided MIM. The encoder refines its capacity to discern relevant semantic features by focusing on regions crucial to the target.

4.3.5 Results on M&Ms Dataset

Finally, we tested PICK on the challenging semi-supervised domain generalization (SSDG) task using the M&Ms dataset. In this group of experiments, competing methods include LLDG , SAML (Liu et al., 2020), DGNet (Liu et al., 2021), vMFNet (Liu et al., 2022), BCP (Bai et al., 2023), and A&D (Wang & Li, 2023). To mitigate the influence of domain shift, we adopted the Fourier Transformation (Xu et al., 2021) as an additional augmentation technique. The results were presented in Table 5. It shows that PICK achieves a 4.09% increase in Dice score compared to the SSDG-based method vMFNet when using 2% labeled data, illustrating the benefits of integrating MIM into SSL. Against the SSL-based method BCP, PICK demonstrates a 6.89% gain in Dice score with 2% labeled data, highlighting the robust representation power of features learned by MIM. Moreover, PICK outperformed the generic SSL method A&D by 1.01% in Dice score with 5% labeled data, further demonstrating the generalizability of our approach.

4.4 Ablation Study

Table 6 presents four combinations of three decoders and their impact on the performance of PICK evaluated on the MSD-Lung Tumor, the NIH-Pancreas, and the Left atrium datasets with 20% labeled data. In Row 1, the primary decoder $f_D(\cdot;\theta_{pri})$ is used as the baseline, learning solely

Table 7 Discussion of the masking strategy

Masking strategy	MSD-lung tumor dataset				NIH-pancreas dataset				Left atrium dataset			
	Dice	Jaccard	ASD	95HD	Dice	Jaccard	ASD	95HD	Dice	Jaccard	ASD	95HD
Random Masking v1	54.69	39.29	3.17	11.29	78.25	64.86	2.14	8.34	89.41	81.00	2.29	7.14
Random Masking v2	55.38	40.26	2.87	11.07	78.31	65.39	2.02	7.98	89.61	81.33	1.98	7.12
Background Masking	53.10	37.85	5.47	14.90	76.13	62.71	2.29	7.15	88.23	79.56	2.32	8.68
Attentive Masking (Ours)	57.65	42.40	2.76	10.85	80.14	67.65	1.55	6.12	90.85	83.37	1.74	5.70

Results are reported on the MSD-Lung Tumor, NIH-Pancreas, and Left Atrium datasets

from labeled data. Row 2 and Row 3 represent variants of PICK where either the MIM decoder $f_D(\cdot;\theta_{mim})$ or the auxiliary decoder $f_D(\cdot;\theta_{aux})$ is excluded, respectively. Row 4 presents the complete PICK framework, incorporating all three decoders. It reveals that, comparing Row 1 to Row 2, the incorporation of the MIM decoder allows effective utilization of unlabeled data through a predict-and-mask strategy, resulting in notable 17.2%, 12.97% and 5.05% improvements in Dice score over the baseline, on the MSD-Lung Tumor, the NIH-Pancreas, and the Left atrium datasets, respectively. Row 3 evaluates the auxiliary decoder's capability to predict on masked but not reconstructed images, exploring direct semantic segmentation on target-masked images. The comparison between Row 3 and Row 4 confirms the importance of reconstructing masked attentive regions, suggesting that explicit reconstruction tasks are more effective. Similarly, the performance gain of Row 4 over Row 2 indicates that relearning from reconstructed images helps resolve conflicts between high-level semantic segmentation and low-level voxel reconstruction tasks. In summary, these experiments highlight the effectiveness of each decoder within the PICK framework, demonstrating their contributions to enhancing segmentation performance.

5 Discussion

5.1 Setting of Parameter λ

The hyper-parameter λ balances the weight between the reconstruction loss and segmentation loss in the PICK framework. To determine the optimal value of λ, we evaluated the performance of PICK using different values of λ, ranging from 0.1 to 1.4 with a step of 0.2, and plotted the results in Fig. 4. It shows that the optimal performance was obtained on the MSD-Lung Tumor, NIH-Pancreas, Left Atrium, AMOS, and M&Ms datasets when setting $lambda$ to 1.0, 1.0, 0.2, 0.2, and 1.2, respectively. Notably, the optimal λ is significantly smaller in the Left Atrium and AMOS datasets compared to the others. This difference can be attributed to the relatively larger size of the segmentation targets in these datasets, which makes the reconstruction loss takes more voxels into consideration and thus necessitates a reduced emphasis on the reconstruction loss during training.

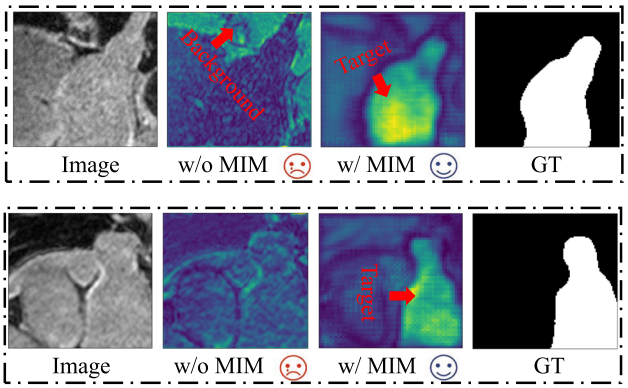

Fig. 5 Two test examples demonstrating the benefits of pseudo-label-guided MIM. Features extracted from the initial convolutional layer of the encoder are shown

5.2 Masking Strategy

The masking strategy plays a pivotal role for training the MIM decoder. For this study, we employed the attentive masking strategy, where the regions to be masked were identified by the pseudo-labels generated by the primary decoder. To validate the effectiveness of this strategy, we compared it with three other masking strategies, including two random masking strategies and the background masking strategy. Random Masking v1 involves randomly masking 60% of the regions with a size of 16×16×16 voxels, while Random Masking v2 applies a single 60% mask over the image. The performance of our PICK using these masking strategies was given in Table 7. Three conclusions can be drawn. First, Random Masking v2 generally outperforms Random Masking v1 across all datasets. For example, on the MSD-Lung dataset, Random Masking v2 achieves higher Dice (55.38% vs. 54.69%) and Jaccard (40.26% vs. 39.29%) scores, while also reducing ASD and 95HD metrics. Similar trends are observed on the NIH-Pancreas and the Left Atrium datasets, indicating that a single larger masked region enhances the model's ability to develop robust feature representations. Second, the Background Masking strategy shows

Table 8 Evaluation of cross-cohort generalizability

Methods	Avg. Dice	ASD	Average Dice of Each Class (AMOS → BTCV) Sp	RK	LK	Ga	Es	Li	St	Ao	IVC	Pa	RAG	LAG
CLD (Lin et al., 2022)	63.19	4.70	*83.7*	82.1	80.6	40.2	50.7	90.0	56.2	81.1	70.5	53.0	36.4	33.8
DHC (Wang & Li, 2023)	66.10	3.93	83.2	84.0	82.8	42.6	54.6	88.9	**63.5**	84.7	73.2	60.1	39.9	35.7
MagicNet (Chen et al., 2023)	*67.22*	*3.54*	83.5	*85.7*	*84.0*	*45.7*	*56.1*	*90.8*	61.2	*85.5*	*74.3*	*60.7*	*40.3*	*38.8*
PICK (Ours)	**69.50**	**2.87**	**87.9**	**88.6**	**84.5**	**46.8**	**59.2**	**91.5**	*62.7*	**87.6**	**77.5**	**63.0**	**42.9**	**41.8**

The models, trained on the AMOS dataset using 10% of the labeled data, were evaluated by making predictions on the unseen BTCV dataset. Notably, the AMOS dataset comprises cases from China, while the BTCV data were collected in the United States, enabling an assessment of the models' ability to generalize across diverse populations

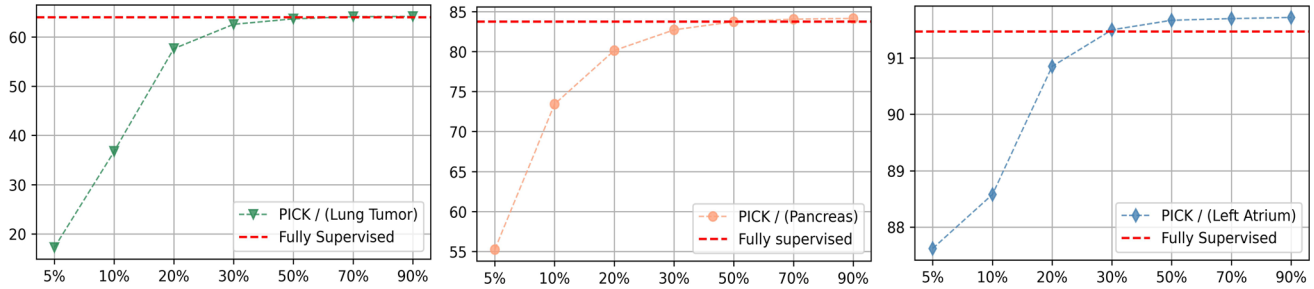

Fig. 6 Dice performance of our PICK model under label percentages of 5%, 10%, 20%, 30%, 50%, 70% and 90%, on the lung tumor dataset, pancreas dataset and left atrium dataset, respectively

the lowest performance among the tested approaches. For instance, on the NIH-Pancreas dataset, it yields the lowest Dice (76.13%) and Jaccard (62.71%) scores, along with higher ASD and 95HD metrics, suggesting that masking less informative regions detrimentally affects model performance. Third, the Attentive Masking strategy proposed in the PICK achieves the best performance across all datasets, with the highest Dice and Jaccard scores and the lowest ASD and 95HD values. For example, on the Left Atrium dataset, Attentive Masking reaches a Dice score of 90.85% and a Jaccard score of 83.37%, significantly outperforming other strategies. This superior performance indicates that our approach, which involves strategically selecting and masking regions based on pseudo-labels, enhances the efficacy of MIM tasks within the SSL framework.

5.3 Advantages Offered by MIM

The performance improvements contributed by the MIM decoder were shown in Table 6 (comparing Row 1 vs. Row 2 and Row 3 vs. Row 4). To further demonstrate the advantages offered by MIM, we randomly selected two test samples and visualized their feature maps obtained from the initial convolutional layer of the encoder with or without the MIM decoder. As shown in Fig. 5, with the MIM decoder, the feature maps more accurately represent target regions, indicating that incorporating pseudo-label-guided MIM into SSL enables the encoder to capture target semantics more effectively through only one time convolution. Moreover, the data reconstructed by MIM serve as augmented views of the source data, which also contribute to enhancing data diversity for the segmentation task.

5.4 Cross-Cohort Evaluation

Table 8 compares the cross-cohort generalizability of PICK with other competitive SSL methods, including MagicNet (Chen et al., 2023), DHC (Wang & Li, 2023), and CLD (Lin et al., 2022). Due to differences in labeled organ categories, we report results only for those annotated in both datasets. Notably, the AMOS data is sourced from China, while the BTCV data (Landman et al., 2015) comes from the United States, highlighting the models' performance across diverse populations. The results show that PICK exhibits more significant performance gains on the BTCV dataset compared to the AMOS test set. Specifically, PICK achieves a 2.28% Dice improvement over MagicNet on the BTCV dataset, compared to a 1.48% improvement on the AMOS test set. This demonstrates PICK's strong generalizability across cohorts, confirming its ability to effectively transfer knowledge from the AMOS dataset (collected in China) to make accurate predictions on the BTCV dataset (collected in the U.S.). We attribute this to PICK's unique design, which embeds unlabeled target information into the encoder, enhancing its ability to capture unseen target semantics. These promising results suggest that PICK is more reliable for real-world clinical applications.

Table 9 Efficiency analysis on the Lung Tumor dataset with an input size of 96×96×96

Method	Dice (%) ↑	95HD (voxel) ↓	Para.(M)	MACs(G)	Training time	Test time / case ↓
CauSSL (Miao et al., 2023)	55.19 ± 0.754	11.97 ± 0.762	18.90	83.88	∼ 4h22 min	∼ 2.96 s
PICK (Ours)	57.65 ± 0.268	10.85 ± 0.436	12.35	47.17	∼ 6h37 min	∼ 2.75s

Metrics include multiply-accumulate operations (MACs), training time, inference time, and parameters (Para.) of the model. (10000 iterations)

5.5 PICK's Results Under Different Label Percentages

Figure 6 presents the Dice performance of PICK under label percentages of 5%, 10%, 20%, 30%, 50%, 70% and 90%, on the MSD-Lung Tumor dataset, NIH-Pancreas dataset and Left Atrium dataset, respectively. It reveals that, on the Left Atrium dataset, just 5% of labeled data achieves a high Dice score of 87.62%, while around 30% labeled data yields performance surpassing the fully supervised model (91.52% vs. 91.47%). This indicates that Left Atrium segmentation can reach optimal results with a relatively small amount of labeled data. In contrast, on the Pancreas dataset, 20% labeled data achieves a Dice score of 80.14%, but approximately 30% is needed to closely match fully supervised performance (82.72% vs. 83.76%). This suggests that pancreas segmentation requires more labeled data than the left atrium to reach peak performance. The Lung Tumor dataset presents the most challenging scenario. Although performance improves steadily as more labeled data is introduced, the model requires around 50% labeled data to approach fully supervised results (63.71% vs. 64.04%). This higher requirement reflects the complexity of lung tumor segmentation, likely due to the variability and irregularity of tumor structures. In summary, PICK demonstrates consistent improvement as more labeled data is provided, converging toward fully supervised performance as the labeled data percentage increases.

5.6 Efficiency Analysis

Table 9 compares the training and inference time of PICK with those of the advanced CauSSL method (Miao et al., 2023), which employs two independent networks for co-training. Despite requiring longer training time, PICK offers quicker inference due to its use of a single decoder for final predictions, whereas CauSSL aggregates the predictions from two networks. This highlights PICK's efficiency during the inference phase.

6 Conclusion

In this paper, we propse a novel approach for SSL in medical image segmentation by integrating the MIM technique. Our method, PICK, intuitively highlights potential targets, facilitating learning from unlabeled data through a predict-and-mask strategy. To bridge the segmentation with the reconstruction task, PICK incorporates a re-learning step on masked and reconstructed images. Extensive experiments conducted across five benchmark datasets in medical imaging demonstrate that PICK outperforms existing methods, showing significant advancements in both SSL and SSDG settings.

7 Limitations and Future Work

The current training process of PICK involves a sequential predict-mask-reconstruct-predict workflow for each sample, which, while effective (with inference requiring only a single prediction using a sliding window approach), can be complex. Future work will explore simplifying PICK's training paradigm through parallel processing approaches to enhance its efficiency.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 62171377 and Grant 92470101, in part by Shenzhen Science and Technology Program under Grants JCYJ20220530161616036, and in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYKFZD06).

Data Availability The data used in this study are all available at (Antonelli et al., 2022; Roth et al., 2015; Xiong et al., 2021; Ji et al., 2022; Campello et al., 2021).

References

Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., & Anderson, P. (2019) Nocaps: Novel object captioning at scale. In: ICCV, pp. 8948–8957.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. (2022). The medical segmentation decathlon. *Nature Communications, 13*(1), 4128.

Bai, Y., Chen, D., Li, Q., Shen, W., & Wang, Y. (2023) Bidirectional copy-paste for semi-supervised medical image segmentation. In: CVPR, pp. 11514–11524.

Campello, V. M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P. M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al. (2021). Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging, 40*(12), 3543–3554.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021) Emerging properties in self-supervised vision transformers. In: ICCV, pp. 9650–9660.

Chen, B., Jiang, J., Wang, X., Wan, P., Wang, J., & Long, M. (2022). *Debiased self-training for semi-supervised learning. NeurIPS, 35*, 32424–32437.

Chen, D., Bai, Y., Shen, W., Li, Q., Yu, L., & Wang, Y. (2023) Magicnet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery. In: CVPR, pp. 23869–23878.

Chen, H., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Savvides, M., & Raj, B. (2022) An embarrassingly simple baseline for imbalanced semi-supervised learning. arXiv preprint arXiv:2211.11086

Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., & Savvides, M. (2022) Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In: ICLR.

Chen, X., Yuan, Y., Zeng, G., & Wang, J. (2021) Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR, pp. 2613–2622.

Chen, Y., Mancini, M., Zhu, X., & Akata, Z. (2022). *Semi-supervised and unsupervised deep visual learning: A survey*. *IEEE Transactions On Pattern Analysis And Machine Intelligence*.

Gal, Y., & Ghahramani, Z. (2016) Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML, pp. 1050–1059 . PMLR.

Gao, S., Zhang, Z., Ma, J., Li, Z., & Zhang, S. (2023) Correlation-aware mutual learning for semi-supervised medical image segmentation. In: MICCAI, pp. 98–108. Springer.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K.Q. (2017) On calibration of modern neural networks. In: ICML, pp. 1321–1330. PMLR.

Guo, L.-Z., Li, & Y.-F. (2022) Class-imbalanced semi-supervised learning with adaptive thresholding. In: ICML, pp. 8082–8094. PMLR.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022) Masked autoencoders are scalable vision learners. In: CVPR, pp. 16000–16009.

Hung, W.-C., Tsai, Y.-H., Liou, Y.-T., Lin, Y.-Y., & Yang, M.-H. (2018) Adversarial learning for semi-supervised semantic segmentation. arXiv preprint arXiv:1802.07934

Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al. (2022). Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *NeurIPS, 35*, 36722–36732.

Jin, Y., Wang, J., & Lin, D. (2022). Semi-supervised semantic segmentation via gentle teaching assistant. *NeurIPS, 35*, 2803–2816.

Kakogeorgiou, I., Gidaris, S., Psomas, B., Avrithis, Y., Bursuc, A., Karantzalos, K., & Komodakis, N. (2022) What to hide from your students: Attention-guided masked image modeling. In: ECCV, pp. 300–318. Springer.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollar, P. (2023) Segment anything. arXiv preprint arXiv:2304.02643

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., & Klein, A. (2015) 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.

Li, H., Wang, Y., Wan, R., Wang, S., Li, T.-Q., & Kot, A. (2020). Domain generalization for medical imaging classification with linear-dependency regularization. *NeurIPS, 33*, 3118–3129.

Li, S., Zhang, C., & He, X. (2020) Shape-aware semi-supervised 3d semantic segmentation for medical images. In: MICCAI, pp. 552–561 . Springer.

Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.W. (2022) Grounded language-image pre-training. In: CVPR, pp. 10965–10975.

Li, Z., Chen, Z., Yang, F., Li, W., Zhu, Y., Zhao, C., Deng, R., Wu, L., Zhao, R., Tang, M., et al. (2021). Mst: Masked self-supervised transformer for visual representation. *NeurIPS, 34*, 13165–13176.

Lin, Y., Yao, H., Li, Z., Zheng, G., & Li, X. (2022) Calibrating label distribution for class-imbalanced barely-supervised knee segmentation. In: MICCAI, pp. 109–118 . Springer.

Liu, Q., Dou, Q., & Heng, P.-A. (2020) Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In: MICCAI, pp. 475–485. Springer.

Liu, X., Thermos, S., O'Neil, A., & Tsaftaris, S.A. (2021). Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation. In: MICCAI, pp. 307–317 Springer.

Liu, X., Thermos, S., Sanchez, P., O'Neil, A.Q., & Tsaftaris, S.A. (2022) vmfnet: Compositionality meets domain-generalised segmentation. In: MICCAI, pp. 704–714 . Springer.

Luo, X., Chen, J., Song, T., & Wang, G. (2021). Semi-supervised medical image segmentation through dual-task consistency. *AAAI, 35*, 8801–8809.

Ma, J., Wang, C., Liu, Y., Lin, L., & Li, G. (2023) Enhanced soft label for semi-supervised semantic segmentation. In: ICCV, pp. 1185–1195.

Miao, J., Chen, C., Liu, F., Wei, H., & Heng, P.-A. (2023) Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In: ICCV, pp. 21426–21437.

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016) V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV, pp. 565–571 . IEEE.

Miyato, T., Maeda, S.-I., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions On Pattern Analysis And Machine Intelligence, 41*(8), 1979–1993.

Na, J., Ha, J.-W., Chang, H.J., Han, D., & Hwang, W. (2023) Switching temporary teachers for semi-supervised semantic segmentation. *NeurIPS*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019) Pytorch: An imperative style, high-performance deep learning library. *NeurIPS***32**.

Peiris, H., Hayat, M., Chen, Z., Egan, G., & Harandi, M. (2023). Uncertainty-guided dual-views for semi-supervised volumetric medical image segmentation. *Nature Machine Intelligenc, 5*(7), 724–738.

Ronneberger, O., Fischer, P., & Brox, T. (2015) U-net: Convolutional networks for biomedical image segmentation. In: MICCAI, pp. 234–241 . Springer.

Roth, H.R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E.B., & Summers, R.M. (2015) Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: MICCAI, pp. 556–564. Springer.

Shen, Z., Cao, P., Yang, H., Liu, X., Yang, J., & Zaiane, O.R. (2023) Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. In: IJCAI.

Tarvainen, A., & Valpola, H. (2017) Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS***30**.

Wang, H., & Li, X. (2023) Dhc: Dual-debiased heterogeneous co-training framework for class-imbalanced semi-supervised medical image segmentation. In: MICCAI, pp. 582–591. Springer.

Wang, H., Li, X. (2023). Towards generic semi-supervised framework for volumetric medical image segmentation. NeurIPS.

Wang, H., Song, K., Fan, J., Wang, Y., Xie, J., & Zhang, Z. (2023) Hard patches mining for masked image modeling. In: CVPR, pp. 10375–10385.

Wang, X., Wu, Z., Lian, L., & Yu, S.X. (2022) Debiased learning from naturally imbalanced pseudo-labels. In: CVPR, pp. 14647–14657.

Wang, Y., Chen, H., Heng, Q., Hou, W., Fan, Y., Wu, Z., Wang, J., Savvides, M., Shinozaki, T., Raj, B., *et al.:* (2022) Freematch: Self-adaptive thresholding for semi-supervised learning. In: ICLR.

Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., & Le, X. (2022) Semi-supervised semantic segmentation using unreliable pseudo-labels. In: CVPR, pp. 4248–4257.

Wang, Y., Xiao, B., Bi, X., Li, W., & Gao, X. (2023) Mcf: Mutual correction framework for semi-supervised medical image segmentation. In: CVPR, pp. 15651–15660.

Wei, C., Sohn, K., Mellina, C., Yuille, A., & Yang, F. (2021) Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: CVPR, pp. 10857–10866.

Wu, F., & Zhuang, X. (2022). Minimizing estimated risks on unlabeled data: A new formulation for semi-supervised medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(5), 6021–6036.

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

Wu, Y., Wu, Z., Wu, Q., Ge, Z., & Cai, J. (2022) Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: MICCAI, pp. 34–43. Springer.

Xie, Y., Zhang, J., Xia, Y., & Shen, C. (2023) Learning from partially labeled data for multi-organ and tumor segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., & Hu, H. (2022) Simmim: A simple framework for masked image modeling. In: CVPR, pp. 9653–9663.

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

Xu, Q., Zhang, R., Zhang, Y., Wang, Y., & Tian, Q. (2021) A fourier-based framework for domain generalization. In: CVPR, pp. 14383–14392.

Yang, L., Qi, L., Feng, L., Zhang, W., & Shi, Y. (2023) Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: CVPR, pp. 7236–7246.

Yu, L., Wang, S., Li, X., Fu, C.-W., & Heng, P.-A. (2019) Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: MICCAI, pp. 605–613 . Springer.

Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., & Yoo, Y. (2019) Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV, pp. 6023–6032.

Zeng, Q., Xie, Y., Lu, Z., Lu, M., & Xia, Y. (2023) Discrepancy matters: Learning from inconsistent decoder features for consistent semi-supervised medical image segmentation. arXiv preprint arXiv:2309.14819

Zeng, Q., Xie, Y., Lu, Z., Lu, M., Wu, Y., & Xia, Y. (2023) Segment together: A versatile paradigm for semi-supervised medical image segmentation. arXiv preprint arXiv:2311.11686

Zeng, Q., Xie, Y., Lu, Z., Lu, M., Zhang, J., Zhou, Y., & Xia, Y. (2024) Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*.

Zeng, Q., Xie, Y., Lu, Z., & Xia, Y. (2023) Pefat: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training. In: CVPR, pp. 15671–15680.