Semi-Mamba-UNet: Pixel-Level Contrastive and Pixel-Level Cross-Supervised Visual Mamba-based UNet for Semi-Supervised Medical Image Segmentation

Chao Ma[a], Ziyang Wang[b,*]

[a] *Mianyang Visual Object Detection and Recognition Engineering Center, Mianyang, China*
[b] *Department of Computer Science, University of Oxford, Oxford, UK*

Abstract

Medical image segmentation is essential in diagnostics, treatment planning, and healthcare, with deep learning offering promising advancements. Notably, Convolutional Neural Network (CNN) excel in capturing local image features, whereas Vision Transformer (ViT) adeptly model long-range dependencies through multi-head self-attention mechanisms. Despite their strengths, both CNN and ViT face challenges in efficiently processing long-range dependencies within medical images, often requiring substantial computational resources. This issue, combined with the high cost and limited availability of expert annotations, poses significant obstacles to achieving precise segmentation. To address these challenges, this paper introduces the Semi-Mamba-UNet, which integrates a purely visual mamba-based U-Shape Encoder-Decoder architecture with a conventional CNN-based UNet into a Semi-Supervised Learning (SSL) framework. This innovative SSL approach leverages both networks to simultaneously generate pseudo labels and cross supervise each other on the pixel level, drawing inspiration from consistency regularization techniques. Furthermore, we introduce a self-supervised pixel-level contrastive learning strategy, employing a pair of projector to further enhance feature learning capabilities especially on unlabeled data. The Semi-Mamba-UNet is comprehensively evaluated on a

*Corresponding Author.
Email address: ziyang.wang@cs.ox.ac.uk (Ziyang Wang)

publicly available MRI cardiac segmentation dataset, comparing against various SSL frameworks with different types of UNet, highlights the superior performance of Semi-Mamba-UNet. The source code of Semi-Mamba-Unet, all baseline methods, corresponding dataset have been made publicly accessible at https://github.com/ziyangwang007/Mamba-UNet.

Keywords: Visual Mamba, UNet, Semi-Supervised Learning, Contrastive Learning, Image Segmentation

---

1. Introduction

Medical image segmentation is essential in enabling precise diagnostics and effective treatment strategies, and deep learning-based networks, particularly those based on the CNN-based UNet architecture, have been extensively investigated [1, 2, 3, 4]. The UNet architecture is with symmetrical encoder-decoder configuration and skip connections at each level. The encoder compresses the input feature map to extract abstract features, which the decoder then uses to reconstruct the image, enhancing the semantic segmentation accuracy. The skip connections are designed to copy and paste feature thus retaining crucial spatial information, further contributing to the network's efficacy. UNet has catalyzed the development of numerous enhancements. For example, U-Net++ [5] introduces a nested UNet structure with deep supervision mechanisms, while Attention UNet [6] incorporates attention gates to bolster the decoders' feature learning capabilities. Moreover, Res-UNet [7] integrates residual learning [8] within its network blocks. Typically, these UNet modifications aim to leverage advanced network constructs such as densenet[9], mobilnet[10], attention mechanism[11] with UNet to improve the feature learning of CNN, thereby addressing the intricate challenges associated with segmenting complex anatomical structures such as CT and MRI [12, 13, 14, 15].

The recent study of multi-head self-attention in sequence-to-sequence tasks has demonstrated the effectiveness of the Transformer network architecture [23]. Image recognition has been proven to benefit from the Vision Transformer,

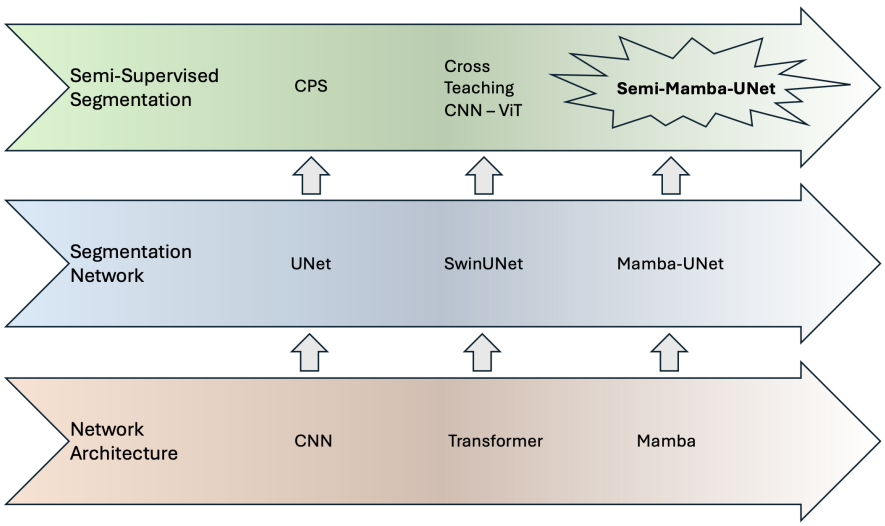

Figure 1: The Development History of Semi-Supervised Learning, Supervised Learning for Medical Image Segmentation, and Network Architecture. Source: CNN [16], Transformer[17], Mamba[18], UNet[1], Swin-UNet[19], Mamba-UNet[20], CPS[21], Cross teaching CNN & ViT[22], and proposed Semi-Mamba-UNet.

which outperform CNN-based networks, especially on large datasets, due to its ability to model long-range dependencies [24]. Several ViT-based networks have also been explored in image segmentation, including SegFormer [25], Segmenter [26], and SETR [27]. In medical image segmentation, most studies related to ViT have also been inspired by UNet, such as TransUNet [28]. This approach explores how Transformer encoders process tokenized image patches from a CNN feature map as the input sequence to extract global contexts, while the decoder upsamples the encoded features. These are then combined with high-resolution CNN feature maps to enable precise localization. SwinUNet [19] further explores the integration of a pure Shift Window-based ViT into a U-shaped architecture, resulting in a pure SwinViT-based UNet. Dense SwinUNet [29] advances this by incorporating deep supervision and densely connected skip connections to enhance segmentation performance. UTNet [30] proposes a computationally efficient self-attention mechanism along with relative position encoding to reduce the complexity of the self-attention operation. Unetr [31] explores a ViT-

based UNet for volumetric medical image segmentation. nnFormer [32], a 3D transformer for volumetric medical image segmentation, not only exploits the combination of interleaved convolution and self-attention operations but also introduces a local and global volume-based self-attention mechanism to learn volume representations.

The efficacy of ViT-based networks, while promising, is contingent upon the availability of extensive labeled datasets, which is challenging to acquire. Weakly-Supervised Learning (WSL) and SSL framework have been investigated [33, 34, 35, 36, 37]. A common approach in these studies involves the integration of UNet with consistency regularization strategies, wherein the network is encouraged to produce consistent outputs under various perturbations. For instance, the Uncertainty-Aware Mean Teacher(UAMT) method employs a UNet architecture within a self-ensembling scheme for feature perturbation and uncertainty estimation[35]. The Cross-Teaching technique extends this concept by leveraging CNN- and ViT-based UNet, enabling collaboration between the two networks through pseudo labels[22]. FixMatch introduces a novel approach by employing both strong and weak data augmentations as forms of data perturbation across networks[38]. Furthermore, multi-view learning expands this cooperative framework to include three networks, promoting mutual learning through co-training [39].

Recent advancements have introduced the novel Mamba architecture, with strength in capturing global contextual information with efficient computational cost, conceptualized by State Space Models (SSMs) [40, 41, 42]. This architecture has been explored in a variety of computer vision tasks, such as Vision Mamba [43], UMamba [44], Segmamba [45], MambaUNet [20], VM-UNet [46], and Weak-Mamba-UNet [47]. In response to the growing need for efficient medical image segmentation, particularly in SSL with limited annotations, this paper introduces the Semi-Mamba-UNet, a novel framework that integrates the Mamba architecture within a pixel-level contrastive, pixel-level cross-supervised learning for semi-supervised medical image segmentation. To the best of our knowledge, this is the first work to explore the Mamba architecture in the med-

ical image segmentation with limited annotations. The development history of UNet and its derivatives in medical image segmentation, and the position of Semi-Mamba-UNet, is depicted in Figure 1. Our contributions are fivefold:

1. Exploration of recent advancement of Visual Mamba as network block into U-Shape Encoder-Decoder style network for medical image segmentation.
2. Integration of a Mamba-based segmentation network with SSL, i.e. a large amount of unlabeled data for network training. For fair evaluation, comparisons are drawn against CNN-based UNet[1] and ViT-based SwinUNet[19] across various SSL frameworks.
3. A pixel-level contrastive learning strategy is introduced with SSL, incorporating a pair of projector to maximize feature learning capabilities using provided both labeled and unlabeled data.
4. A pixel-level cross-supervised learning is introduced with SSL. The network trained with the help of the other network via pseudo labeling, thereby extending the utility of unlabeled data in network training.
5. Semi-Mamba-UNet is validated with a public benchmark dataset, demonstrating State-of-the-Art performance. The source code of Semi-Mamba-UNet and all baseline methods are made public available.

2. Methodology

The framework of Semi-Mamba-UNet is illustrated in Figure 2. As shown in Figure, $(\boldsymbol{X}_{\mathrm{l}}, \boldsymbol{Y}_{\mathrm{gt}}) \in \mathbf{L}$ denotes as the labeled training data set. $(\boldsymbol{X}_{\mathrm{u}}) \in \mathbf{U}$ denotes as the unlabeled training data set. And $(\boldsymbol{X}_{\mathrm{t}}, \boldsymbol{Y}_{\mathrm{t}}) \in \mathbf{T}$ denote the labeled testing data set. $\boldsymbol{X} \in \mathbb{R}^{h \times w}$ represents a 2D grayscale image and the size is h high and w width. $\boldsymbol{Y}_{\mathrm{l}}, \boldsymbol{Y}_{\mathrm{t}} \in \mathbb{N}_4^{h \times w}$ represents a 4-class labeled segmentation mask with pixel values ranging from 0 to 3, indicating as right ventricle (RVC), left ventricle (LVC) and myocardium (MYO). The predicted segmentation mask by a segmentation network given $\boldsymbol{X}$ as $Y_{\mathrm{p}} = f(X; \theta)$ with the θ as parameters. The Mamba-based UNet and the UNet are denoted as $f_1(\theta)$ and $f_2(\theta)$. The

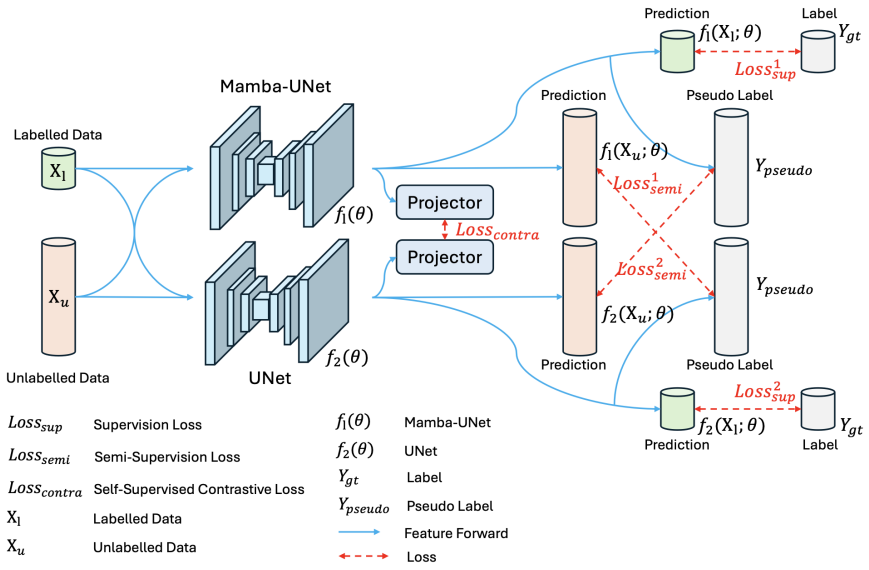

Figure 2: Semi-Mamba-UNet: The Framework of Pixel-Level Contrastive Cross-Supervised Visual Mamba-based UNet for Semi-Supervised Medical Image Segmentation.

prediction of a network can be considered as pseudo label to expand the unlabeled dataset as $(\boldsymbol{X}_{\mathrm{u}}, \boldsymbol{Y}_{\mathrm{pseudo}}) \in \mathbf{U}$ to train the other network. A pair of projectors $p(\cdot)$ is introduced to each network to extract representation features of training set for contrastive learning purposes. The overall losses are categorized as supervision loss $\mathcal{L}_{\mathrm{sup}}$, semi-supervised loss $\mathcal{L}_{\mathrm{semi}}$, and self-supervised contrastive loss $\mathcal{L}_{\mathrm{contra}}$. The evaluation is conducted by measuring the difference between $(\boldsymbol{Y}_{\mathrm{p}}, \boldsymbol{Y}_{\mathrm{t}})$ on the test set. The overall training objective is to update the parameters of network θ thus minimizing the total loss $\mathcal{L}_{total}$, illustrated as:

$$\mathcal{L}_{total} = \underbrace{\mathcal{L}^{1}_{\mathrm{sup}} + \mathcal{L}^{2}_{\mathrm{sup}}}_{sup} + \underbrace{\mathcal{L}^{1}_{\mathrm{semi}} + \mathcal{L}^{2}_{\mathrm{semi}}}_{semi} + \underbrace{\mathcal{L}_{\mathrm{contra}}}_{self} \tag{1}$$

All mathematical symbols are denoted in the Figure 2 accordingly, and $\mathcal{L}$ is highlighted as red dash line, where *sup* is the supervision loss for $f_1(\theta)$ and $f_2(\theta)$ based on labeled training set. The $\mathcal{L}_{\mathrm{sup}}$ is designed with the combination of Dice-Coefficient-based (Dice) and Cross-Entropy-based (CE) loss as following,

$$\mathcal{L}_{\text{sup}}^{1} = \text{CE}\big(\text{softmax}(f_1(\boldsymbol{X}_u;\theta),\, \boldsymbol{Y}_{\text{gt}})\big) + \text{Dice}\big(\text{softmax}(f_1(\boldsymbol{X}_u;\theta),\, \boldsymbol{Y}_{\text{gt}})\big) \quad (2)$$

$$\mathcal{L}_{\text{sup}}^{2} = \text{CE}\big(\text{softmax}(f_2(\boldsymbol{X}_u;\theta),\, \boldsymbol{Y}_{\text{gt}})\big) + \text{Dice}\big(\text{softmax}(f_2(\boldsymbol{X}_u;\theta),\, \boldsymbol{Y}_{\text{gt}})\big) \quad (3)$$

$\mathcal{L}_{\text{semi}}$ is the semi-supervision loss for $f_1(\theta)$ and $f_2(\theta)$ based on unlabeled training set. A prediction of a network is considered as the pseudo label $\boldsymbol{Y}_{\text{pseudo}}$ to extend $\boldsymbol{X}_{\text{u}}$ to retrain the other networks. $\mathcal{L}_{\text{contra}}$ is the self-supervised contrastive learning loss, and we proposes a projector pair to extract features between the prediction of two networks. The details of Mamba-UNet, pixel-level cross-supervised learning with $\mathcal{L}_{\text{semi}}$, and pixel-level contrastive learning with $\mathcal{L}_{\text{contra}}$ are discussed in the following sections.

2.1. Mamba-UNet

The UNet architecture, as depicted in Figure 3, represents a novel adaptation of the conventional encoder-decoder style segmentation network with various types of network blocks for medical image analysis. To ensure a fair comparison, the proposed utilization of Mamba-UNetis developed against the original UNet [1] and the Swin UNet [19]. Each of these networks adheres to the U-shaped encoder-decoder configuration. Specifically, the UNet employs 2-layer CNN with the size of 3×3[1], the Swin-UNet utilizes 2 Swin Transformer blocks[19], and the Mamba-UNet integrates 2 Visual Mamba blocks, which is part of our past work in [20]. This distinction in block composition is pivotal, as it directly influences the networks' ability to process and interpret the intricate details present in medical images.

Specifically, the conventional SSMs as a linear time-invariant system function to map $x(t) \in \mathbb{R} \mapsto y(t) \in \mathbb{R}$ through a hidden state $h(t) \in \mathbb{R}^N$, given $A \in \mathbb{C}^{N\times N}$ as the evolution parameter, $B, C \in \mathbb{C}^N$ as the projection parameters for a state size N, and skip connection $D \in \mathbb{C}^1$. The model can be formulated as linear ordinary differential equations (ODEs) in Eq 4,

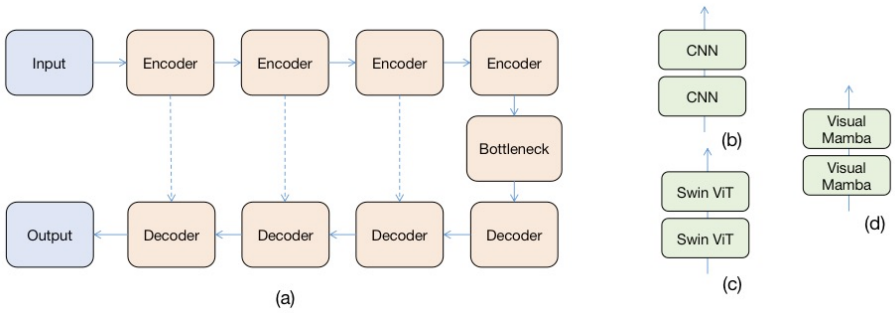

Figure 3: The Segmentation Backbone Network in This Study. (a) Encoder-Decoder Style Segmentation Network. (b) The 2-Layer CNN-based Network Block of UNet. (c) The 2-Layer Swin ViT-based Network Block of Swin-UNet. (d) The 2-Layer Visual Mamba-based Network Block of Mamba-UNet.

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t) + Dx(t). \end{aligned} \tag{4}$$

The discrete version of this linear model can be transformed by zero-order hold given a timescale parameter $\Delta \in \mathbb{R}^D$.

$$\begin{aligned} h_t &= \overline{A}h_{k-2} + \overline{B}x_k \\ y_t &= Ch_k + \overline{D}x_k \\ \overline{A} &= e^{\Delta A} \\ \overline{B} &= (e^{\Delta A} - I)A^{-1}B \\ \overline{C} &= C \end{aligned} \tag{5}$$

where $B, C \in \mathbb{R}^{D\times N}$. The approximation of $\overline{B}$ refined using first-order Taylor series $\overline{B} = \left(e^{\Delta A} - I\right) A^{-1}B \approx (\Delta A)\,(\Delta A)^{-1}\,\Delta B = \Delta B$. The Visual Mamba further introduce Cross-Scan Module (CSM) then integrate convolutional operations into the block, which is detailed in [41, 18]. The Mamba-UNet, with its Visual Mamba blocks, aims to capitalize on the efficiency and effectiveness of Mamba models in capturing and processing complex spatial and contextual information, thereby enhancing segmentation performance.

2.2. Pixel-Level Cross-Supervised Learning

Inspired by the principle of consistency regularization and multi-view learning, such as the Cross Pseudo Supervision [21], where two independently initialized networks generate and exchange pseudo labels for mutual supervision, this work extends the concept to leverage the complementary strengths of distinct architectures. The methodology of Cross Teaching between CNN and ViT [48] further explores on the mutual benefits derived from the collaboration between two different network architectures. Similarly, FixMatch [38] advocates for the application of two distinct data augmentations across two networks, with one network acting as a supervisor for the other through the use of augmented data. In the Semi-Mamba-UNet, we introduce simple yet efficient cross-supervised learning strategy enabling Mamba-UNet and UNet directly help each other, and $\mathcal{L}_{\text{semi}}$ is illustrated as,

$$\mathcal{L}^{1}_{\text{semi}} = \text{CE}\big(\text{argmax}(f_1(\boldsymbol{X}_u;\theta), f_2(\boldsymbol{X}_u;\theta))\big) + \text{Dice}\big(\text{argmax}(f_1(\boldsymbol{X}_u;\theta), f_2(\boldsymbol{X}_u;\theta))\big) \tag{6}$$

$$\mathcal{L}^{2}_{\text{semi}} = \text{CE}\big(\text{argmax}(f_2(\boldsymbol{X}_u;\theta), f_1(\boldsymbol{X}_u;\theta))\big) + \text{Dice}\big(\text{argmax}(f_2(\boldsymbol{X}_u;\theta), f_1(\boldsymbol{X}_u;\theta))\big) \tag{7}$$

2.3. Pixel-Level Contrastive Learning

Contrastive learning has been recognized as a potent paradigm for the derivation of robust and discriminative features, representing a significant stride in the realm of self-supervised learning [49]. The key idea behind contrastive learning is that positive and negative samples are discriminative. Positive and negative samples are initially constructed based on prior knowledge and map them to a potential feature embedding space. A metric function is then utilized to encourage the network to bring positive samples closer together and distance the positive from the negative samples. Contrastive learning has demonstrated remarkable efficacy across a spectrum of applications [50, 51, 52].

The application of contrastive learning of medical image analysis addresses the perennial challenges posed by sparse annotations and augments the capacity

for feature extraction, culminating in enhanced model performance [53, 54, 55, 56, 57]. Most of latest work constructs positive and negative samples by specific tracks such as applying different perturbations to the same sample, and encoding the same sample by different types of encodes, focusing on the variability of sample attributes. In order to better understand the variability of pixels between samples, a novel pixel-level contrsitive learning is proposed.

Considering the small size of cardiac, while a large amount of pixels belongs to background, the consistency of these pixels is insignificant for network training. An adaptive average pooling as a projector to filter the unwanted background pixels and highlight the target region's representational ability in the image is proposed. Further, we apply L2 regularisation in the channel dimension to sparse the features to improve the model's resistance to perturbation. Inter-network consistency by computing the mean square error between features is utilized further. In the proposed SSL framework, we utilize a projector pair to Mamba-UNet, and UNet simultaneously. This configuration facilitates the extraction of pixel-level feature representations, which subsequently serve as the basis for computing image similarity within the defined feature space. The similarity assessment is conducted according to [58], formalized as,

$$\mathcal{L}_{\text{contra}} = \frac{\sum \|(G(F_\theta(X_L \cup X_U)), G(F_\theta(X_L \cup X_U)))\|_2^2}{N} \tag{8}$$

where F_θ is a predictor which has the same AdaptiveAvgPool as the projector, G is l_2 regularization along the channel axis and N is the number of input data. X_L and X_U represent labeled and unlabeled data, respectively, while $\cup$ represents union with a mathematical symbol. To effectively leverage data set for network training, we further assume labeled data as unlabeled data to expand the dataset (i.e. $\cup$) in the process of consistency regularization of unlabeled set to boost the performance, which is different with conventional SSL strategies.

3. Experiments and Results

Datasets: The efficacy of Semi-Mamba-UNet, alongside various baseline methodologies, was assessed using a publicly available MRI cardiac segmentation dataset, namely the ACDC dataset from the MICCAI 2017 Challenge [59]. This dataset encompasses imaging data from 100 patients, providing a comprehensive basis for evaluation. To comply with the input requirements of SwinUNet [19], all images are resized to 224×224 pixels. The dataset was partitioned such that 20% constituted the testing set, with the remaining 80% allocated for training and validation purposes. The experimental setup was designed to simulate scenarios where only 5% and 10% of the training set were available as labeled data. The labeled & unlabeled data is randomly selected from ACDC only once, and utilize for Semi-Mamba-UNet and all baseline methods. There is no overlap between labeled training set, unlabeled training set, validation set, and test set.

Implementation Details: The development environment for our experiments was Ubuntu 20.04, utilizing PyTorch. The computational hardware included an Nvidia GeForce RTX 3090 GPU and an Intel Core i9-10900K CPU. The average runtime for the experiments was approximately 5-8 hours. The dataset is designed to 2D image segmentation tasks. Training of the Semi-Mamba-UNet encompassed 30,000 iterations with a batch size of 16. The Stochastic Gradient Descent (SGD) optimizer was employed, featuring a learning rate of 0.01, momentum of 0.9, and a weight decay of 0.0001. Evaluation was conducted on the validation set at every 200 iterations, with the network's weights being preserved only if the validation performance surpassed previous best network.

Baseline Segmentation Networks: The framework of Semi-Mamba-UNet is depicted in Figure 2 with two segmentation backbone networks. To ensure equitable comparisons, we also employed the CNN-based UNet [1] and the Swin ViT-based SwinUNet [19] as segmentation backbone networks for different SSL frameworks. This selection was motivated by the architectural similarities these

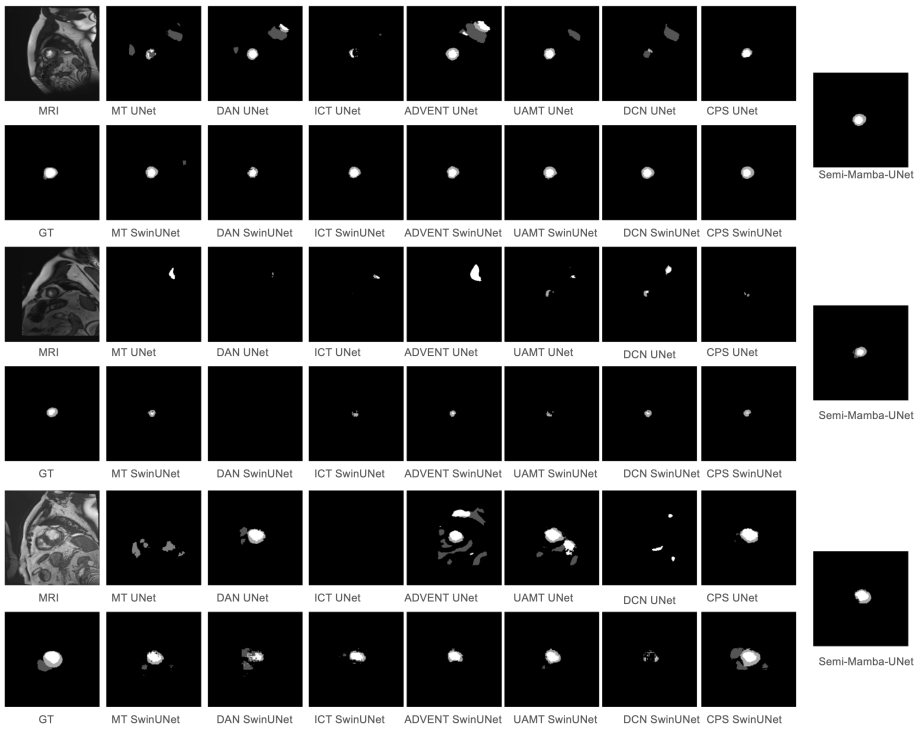

Figure 4: Three Randomly Selected Example MRI Images, Ground Truth, and Corresponding Segmentation Results of all Baseline Methods and Semi-Mamba-UNet when 5% of Data are Assumed as Labeled Data.

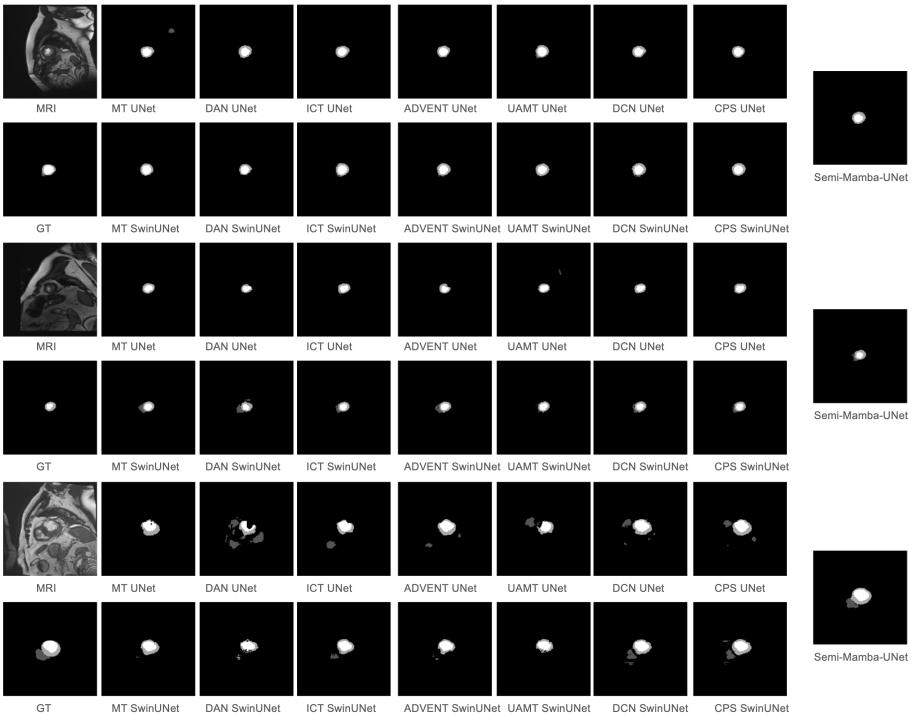

Figure 5: Three Randomly Selected Example MRI Images, Ground Truth, and Corresponding Segmentation Results of all Baseline Methods and Semi-Mamba-UNet when 10% of Data are Assumed as Labeled Data.

networks share with our proposed framework, thereby providing a consistent basis for evaluating the performance enhancements introduced by the Semi-Mamba-UNet.

Baseline SSL Frameworks: The SSL baseline frameworks evaluated includes Mean Teacher (MT) [60], Deep Adversarial Network (DAN) [61], Interpolation Consistency Training (ICT) [62], Adversarial Entropy Minimization (ADVENT) [63], Uncertainty Aware Mean Teacher (UAMT) [35], and Deep Co-Training (DCN) [64]. Both SwinUNet [19] and UNet [1] were employed as the segmentation backbone networks across all above SSL frameworks.

Table 1: Direct Comparison of Semi-supervised Frameworks on MRI Cardiac Test Set when **5%** of Data is Assumed as Labeled Data

SSL Framework+Network	Dice↑	Acc↑	Pre↑	Sen↑	Spe↑	HD↓	ASD↓
MT[60] + SwinUNet	0.7506	0.9910	0.7918	0.7178	0.9394	10.4621	3.5301
DAN[61] + SwinUNet	0.7252	0.9901	0.7695	0.6903	0.9337	12.9800	4.3823
ICT[62] + SwinUNet	0.7504	0.9910	0.7923	0.7180	0.9392	9.8026	3.1055
ADVENT[63]+ SwinUNet	0.7489	0.9910	0.7964	0.7128	0.9373	11.1535	3.0907
UAMT[35] + SwinUNet[65]	0.7442	0.9909	0.7902	0.7108	0.9393	10.2955	2.8222
DCN[64] + SwinUNet	0.7603	0.9914	0.8118	0.7207	0.9376	10.1783	3.1478
CPS[21] + SwinUNet	0.7901	0.9919	0.8162	0.7730	0.9533	8.4888	2.1833
MT[60] + UNet	0.7256	0.9885	0.8217	0.6670	0.9044	24.0480	9.7662
DAN[61] + UNet	0.7657	0.9905	0.8296	0.7152	0.9199	21.1226	7.3434
ICT[62] + UNet	0.7490	0.9906	0.8827	0.6633	0.9013	11.2109	4.5181
ADVENT[63]+ UNet	0.6656	0.9833	0.6487	0.6900	0.9190	42.8011	16.6207
UAMT[35] + UNet	0.7472	0.9901	0.8164	0.6943	0.9168	21.7492	7.7489
DCN[64] + UNet	0.7312	0.9894	0.8316	0.6626	0.9022	24.6607	10.1996
CPS[21] + UNet	0.7699	0.9912	0.9084	0.6829	0.9039	**6.1406**	**1.1477**
Semi-Mamba-UNet	**0.8386**	**0.9936**	**0.8861**	**0.7992**	**0.9483**	6.2139	1.6406

Evaluation Metrics: To assess the performance of Semi-Mamba-UNet against other SSL baseline methods, comprehensive evaluation metrics are employed. Similarity measures include Dice Coefficient(Dice), Accuracy(Acc), Precision(Pre), Sensitivity(Sen), and Specificity(Spe).

Table 2: Direct Comparison of Semi-supervised Frameworks on MRI Cardiac Test Set when **10%** of Data is Assumed as Labeled Data

Framework+Network	Dice↑	Acc↑	Pre↑	Sen↑	Spe↑	HD↓	ASD↓
MT[60] + SwinUNet	0.8678	0.9949	0.8700	0.8670	0.9745	7.3576	2.1834
DAN[61] + SwinUNet	0.8288	0.9936	0.8261	0.8375	0.9721	9.9132	2.7309
ICT[62] + SwinUNet	0.8621	0.9947	0.8624	0.8632	0.9746	8.7211	2.5562
ADVENT[63]+ SwinUNet	0.8669	0.9949	0.8688	0.8660	0.9743	7.1383	2.2608
UAMT[35] + SwinUNet[65]	0.8701	0.9950	0.8721	0.8697	0.9754	6.7226	2.0975
DCN[64] + SwinUNet	0.8608	0.9946	0.8511	0.8724	0.9777	8.8474	2.6705
CPS[21] + SwinUNet	0.8933	0.9957	0.8846	0.9032	0.9821	5.5661	1.6418
MT[60] + UNet	0.8781	0.9949	0.8836	0.8735	0.9690	10.9691	3.3246
DAN[61] + UNet	0.8766	0.9948	0.8814	0.8727	0.9700	8.6977	2.4750
ICT[62] + UNet	0.8879	0.9953	0.8996	0.8779	0.9696	6.7011	1.9696
ADVENT[63]+ UNet	0.8777	0.9949	0.8877	0.8703	0.9674	11.0979	2.9367
UAMT[35] + UNet	0.8798	0.9949	0.8778	0.8821	0.9726	10.2134	3.1926
DCN[64] + UNet	0.8831	0.9952	0.8897	0.8785	0.9706	8.6978	2.7026
CPS[21] + UNet	0.8933	0.9956	0.8965	0.8912	0.9749	7.8319	2.2767
Semi-Mamba-UNet	**0.9114**	**0.9964**	**0.9088**	**0.9146**	**0.9821**	**3.9124**	**1.1698**

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} \tag{9}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{12}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{13}$$

where, TP, FP, TN, and FN represent the number of True Positives, False Positives, True Negatives, and False Negatives prediction on each pixel. These metrics provide a comprehensive assessment of the network's performance across various aspects, which is the higher the better, denoted as ↑.

Additionally, difference measures, where lower values are preferable↓, consist of the 95% Hausdorff Distance (HD) and Average Surface Distance (ASD). Given that the dataset involves 4-class segmentation tasks, the mean values of

Table 3: Ablation Studies on Contributions Segmentation Backbone Network with the Same SSL Framework.

Ratio	Network	Dice↑	Acc↑	Pre↑	Sen↑	Spe↑	HD↓	ASD↓
5%	2 × SwinUNet	0.7878	0.9918	0.8066	0.7795	**0.9577**	9.0240	2.3592
5%	2 × Mamba-UNet	0.8025	0.9924	0.8623	0.7558	0.9379	7.3952	2.1257
5%	UNet + SwinUNet	0.8292	0.9933	0.8591	**0.8052**	0.9557	**5.7014**	1.7237
5%	UNet + Mamba-UNet	**0.8386**	**0.9936**	**0.8861**	0.7992	0.9483	<u>6.2139</u>	**1.6406**
10%	2 × SwinUNet	0.8899	0.9955	0.8784	0.9031	0.9823	5.9222	1.6960
10%	2 × Mamba-UNet	0.9006	0.9959	0.8913	0.9109	**0.9826**	6.7631	1.8349
10%	UNet + SwinUNet	0.9105	0.9963	0.9057	**0.9161**	**0.9826**	5.4172	1.4506
10%	UNet + Mamba-UNet	**0.9114**	**0.9964**	**0.9088**	<u>0.9146</u>	0.9821	**3.9124**	**1.1698**

these metrics across all classes are reported.

Qualitative Results: Figure 4 and Figure 5 illustrates randomly selected sample raw MRI scan, ground truth, and corresponding prediction of all SSL baseline frameworks with several types of UNet including Semi-Mamba-UNet under different data situations (5% and 10% of training set as labeled data).

Quantitative Results: The performance of Semi-Mamba-UNet in direct comparison with other SSL methods is quantitatively detailed in Table 1, encompassing both similarity and difference measures, under the condition where the ratio of labeled to total data is set at 5%. Table 2 extends this comparison to a scenario where the labeled data ratio is increased to 10%. In both tables, the highest-performing metrics are highlighted in **bold**, and the second best of Semi-Mamba-UNet is with <u>Underline</u>.

Evaluation on Each Image of Test Set: Except for reporting mean value of each evaluation metrics in Table For a comprehensive evaluation with Table 1, 2, we also report segmentation result on the each of image on the test set. A histogram is briefly sketched in Figure 6 (a), and (b) where the X-axis is the IoU, and the Y-axis is the number of predicted images on test set with the corresponding IoU score, demonstrating Semi-Mamba-UNet is more likely

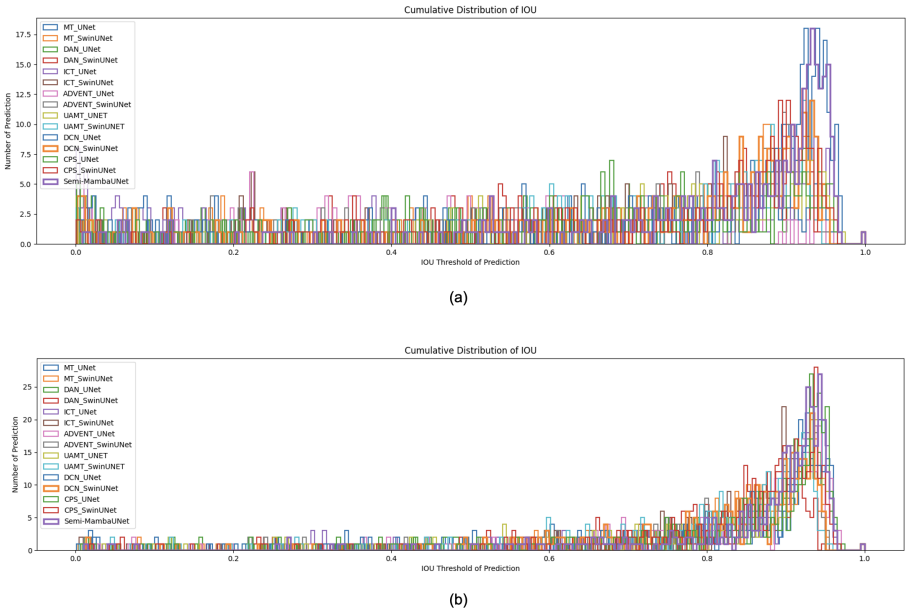

Figure 6: The Distribution of Segmentation Results According to IoU. (a) When 5% of Data as Labeled Data. (b) When 10% of Data as Labeled Data.

to predict images with high IoU scores against other baseline methods.

Ablation Study: The ablation studies presented in Table 3 illustrates the contributions of the proposed SSL framework with other advanced segmentation backbone networks, i.e. UNet, SwinUNet, and Mamba-UNet. The studies are also conducted when the labeled data constitutes 5% and 10% of the total dataset. The ablation study further demonstrates the effective of proposed Semi-Mamba-UNet.

4. Conclusion

In this study, we investigated the integration of Visual Mamba within the UNet architecture with a semi-supervised fashion for medical image segmentation. An advanced semi-supervised learning strategy, combining pixel-level cross-supervision with pixel-level contrastive learning, to harness the full potential of the Visual Mamba. Our extensive experimental evaluations demonstrate

the effectiveness of the Semi-Mamba-UNet. In the future, we aim to extend our research to encompass volumetric data segmentation and further refine our methods within the scope of limited-supervised learning scenarios, continuing to leverage the unique capabilities of the Visual Mamba.

C.M.: conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, visualization, writing—original draft, writing—review and editing. Z.W.: conceptualization, formal analysis, investigation, methodology, project administration, supervision, validation, writing—original draft, review and editing. All authors have read and agreed to the published version of the manuscript.

References

[1] O. Ronneberger, et al., U-Net: Convolutional networks for biomedical image segmentation, in: Int Conf Med Im Comp & Comp-Assisted Intervention, Springer, 2015, pp. 234–241.

[2] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), IEEE, 2016, pp. 565–571.

[3] N. Ibtehaz, M. S. Rahman, Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation, Neural networks 121 (2020) 74–87.

[4] Z. Wang, I. Voiculescu, Quadruple augmented pyramid network for multi-class covid-19 segmentation via ct, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2021, pp. 2956–2959.

[5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.

[6] O. Oktay, et al., Attention U-Net: Learning where to look for the pancreas, Int Conf Medical Imaging with Deep Learning (2018).

[7] F. I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data, ISPRS Journal of Photogrammetry and Remote Sensing 162 (2020) 94–114.

[8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[9] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).

[11] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[12] Y. Zhang, L. Yuan, Y. Wang, J. Zhang, Sau-net: efficient 3d spine mri segmentation using inter-slice attention, in: Medical Imaging With Deep Learning, PMLR, 2020, pp. 903–913.

[13] Z. Wang, Z. Zhang, I. Voiculescu, Rar-u-net: a residual encoder to attention decoder by residual connections framework for spine segmentation under noisy labels, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 21–25.

[14] A. Chaurasia, E. Culurciello, Linknet: Exploiting encoder representations for efficient semantic segmentation, in: 2017 IEEE Visual Communications and Image Processing (VCIP), IEEE, 2017, pp. 1–4.

[15] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, P.-A. Heng, H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes, IEEE transactions on medical imaging 37 (12) (2018) 2663–2674.

[16] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[17] Z. Liu, Y. Lin, et al., Swin transformer: Hierarchical vision transformer using shifted windows, arXiv preprint arXiv:2103.14030 (2021).

[18] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, Y. Liu, Vmamba: Visual state space model, arXiv preprint arXiv:2401.10166 (2024).

[19] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European conference on computer vision, Springer, 2022, pp. 205–218.

[20] Z. Wang, et al., Mamba-unet: Unet-like pure visual mamba for medical image segmentation, arXiv preprint arXiv:2402.05079 (2024).

[21] X. Chen, Y. Yuan, G. Zeng, J. Wang, Semi-supervised semantic segmentation with cross pseudo supervision, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2613–2622.

[22] X. Luo, et al., Semi-supervised medical image segmentation via cross teaching between cnn and transformer, arXiv preprint arXiv:2112.04894 (2021).

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[25] E. Xie, et al., Segformer: Simple and efficient design for semantic segmentation with transformers, Advances in Neural Information Processing Systems 34 (2021) 12077–12090.

[26] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: Transformer for semantic segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 7262–7272.

[27] S. Zheng, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 6881–6890.

[28] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, arXiv preprint arXiv:2102.04306 (2021).

[29] Z. Wang, M. Su, J.-Q. Zheng, Y. Liu, Densely connected swin-unet for multiscale information aggregation in medical image segmentation, in: 2023 IEEE International Conference on Image Processing (ICIP), IEEE, 2023, pp. 940–944.

[30] Y. Gao, M. Zhou, D. N. Metaxas, Utnet: a hybrid transformer architecture for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, Springer, 2021, pp. 61–71.

[31] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, D. Xu, Unetr: Transformers for 3d medical image segmenta-

tion, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 574–584.

[32] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, Y. Yu, nnformer: Volumetric medical image segmentation via a 3d transformer, IEEE Transactions on Image Processing (2023).

[33] X. Luo, M. Hu, W. Liao, S. Zhai, T. Song, G. Wang, S. Zhang, Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 528–538.

[34] Z. Wang, I. Voiculescu, Exigent examiner and mean teacher: An advanced 3d cnn-based semi-supervised brain tumor segmentation framework, in: Workshop on Medical Image Learning with Limited and Noisy Data, Springer, 2023, pp. 181–190.

[35] L. Yu, S. Wang, X. Li, C.-W. Fu, P.-A. Heng, Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 605–613.

[36] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, P.-A. Heng, Transformation-consistent self-ensembling model for semisupervised medical image segmentation, IEEE Transactions on Neural Networks and Learning Systems 32 (2) (2020) 523–534.

[37] Z. Wang, I. Voiculescu, Weakly supervised medical image segmentation through dense combinations of dense pseudo-labels, in: MICCAI Workshop on Data Engineering in Medical Imaging, Springer, 2023, pp. 1–10.

[38] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, C.-L. Li, Fixmatch: Simplifying semi-supervised learn-

ing with consistency and confidence, Advances in neural information processing systems 33 (2020) 596–608.

[39] Y. Xia, F. Liu, D. Yang, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, H. Roth, 3d semi-supervised learning with uncertainty-aware multi-view co-training, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 3646–3655.

[40] J. Wang, W. Zhu, P. Wang, X. Yu, L. Liu, M. Omar, R. Hamid, Selective structured state-spaces for long-form video understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6387–6397.

[41] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, arXiv preprint arXiv:2312.00752 (2023).

[42] A. Gu, Modeling sequences with structured state spaces, Ph.D. thesis, Stanford University (2023).

[43] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision mamba: Efficient visual representation learning with bidirectional state space model, arXiv preprint arXiv:2401.09417 (2024).

[44] J. Ma, F. Li, B. Wang, U-mamba: Enhancing long-range dependency for biomedical image segmentation, arXiv preprint arXiv:2401.04722 (2024).

[45] Z. Xing, T. Ye, Y. Yang, G. Liu, L. Zhu, Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation, arXiv preprint arXiv:2401.13560 (2024).

[46] J. Ruan, S. Xiang, Vm-unet: Vision mamba unet for medical image segmentation, arXiv preprint arXiv:2402.02491 (2024).

[47] Z. Wang, C. Ma, Weak-mamba-unet: Visual mamba makes cnn and vit work better for scribble-based medical image segmentation, arXiv preprint arXiv:2402.10887 (2024).

[48] X. Luo, M. Hu, T. Song, G. Wang, S. Zhang, Semi-supervised medical image segmentation via cross teaching between cnn and transformer, in: International conference on medical imaging with deep learning, PMLR, 2022, pp. 820–833.

[49] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[50] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.

[51] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.

[52] G. Kang, et al., Contrastive adaptation network for unsupervised domain adaptation, in: CVPR, 2019, pp. 4893–4902.

[53] K. Chaitanya, et al., Contrastive learning of global and local features for medical image segmentation with limited annotations, NIPS (2020).

[54] X. Hu, D. Zeng, X. Xu, Y. Shi, Semi-supervised contrastive learning for label-efficient medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer, 2021, pp. 481–490.

[55] C. You, et al., Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation, IEEE TMI (2022).

[56] Z. Wang, C. Ma, Dual-contrastive dual-consistency dual-transformer: A semi-supervised approach to medical image segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 870–879.

[57] W. Shi, Z. Zhou, B. H. Letcher, N. Hitt, Y. Kanno, R. Futamura, O. Kishida, K. Morita, S. Li, Aging contrast: A contrastive learning framework for fish re-identification across seasons and years, in: Australasian Joint Conference on Artificial Intelligence, Springer Nature Singapore Singapore, 2023, pp. 252–264.

[58] Y. Xie, J. Zhang, Z. Liao, Y. Xia, C. Shen, Pgl: Prior-guided local self-supervised learning for 3d medical image segmentation, arXiv preprint arXiv:2011.12640 (2020).

[59] O. Bernard, 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) (2018) 2514–2525.

[60] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 1195–1204.

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

[62] V. Verma, A. Lamb, J. Kannala, Y. Bengio, D. Lopez-Paz, Interpolation consistency training for semi-supervised learning, in: International Joint Conference on Artificial Intelligence, 2019, pp. 3635–3641.

[63] T.-H. Vu, H. Jain, M. Bucher, M. Cord, P. Pérez, Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2517–2526.

[64] S. Qiao, W. Shen, Z. Zhang, B. Wang, A. Yuille, Deep co-training for semi-

supervised image recognition, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 135–152.

[65] Z. Wang, An uncertainty-aware transformer for mri cardiac semantic segmentation via mean teachers, Annual Conference on Medical Image Understanding and Analysis (2022).