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

Ziyang Wang[1⋆] and Chao Ma[2]

[1] Department of Computer Science, University of Oxford, UK
[2] Mianyang Visual Object Detection and Recognition Engineering Center, China
ziyang.wang@cs.ox.ac.uk
https://github.com/ziyangwang007/Mamba-UNet

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 visual mamba-based UNet architecture with a conventional UNet into a semi-supervised learning (SSL) framework. This innovative SSL approach leverages dual networks to jointly generate pseudo labels and cross supervise each other, drawing inspiration from consistency regularization techniques. Furthermore, we introduce a self-supervised pixel-level contrastive learning strategy, employing a projector pair to further enhance feature learning capabilities. Our comprehensive evaluation on a publicly available MRI cardiac segmentation dataset, comparing against various SSL frameworks with different UNet-based segmentation networks, highlights the superior performance of Semi-Mamba-UNet. The source code has been made publicly accessible.

Keywords: Medical Image Segmentation, Visual Mamba, Semi-Supervised Learning.

1 Introduction

Medical image segmentation is essential in enabling precise diagnostics and effective treatment strategies, and deep learning-based networks, particularly those

⋆ Ziyang Wang and Chao Ma contributed equally to this work.

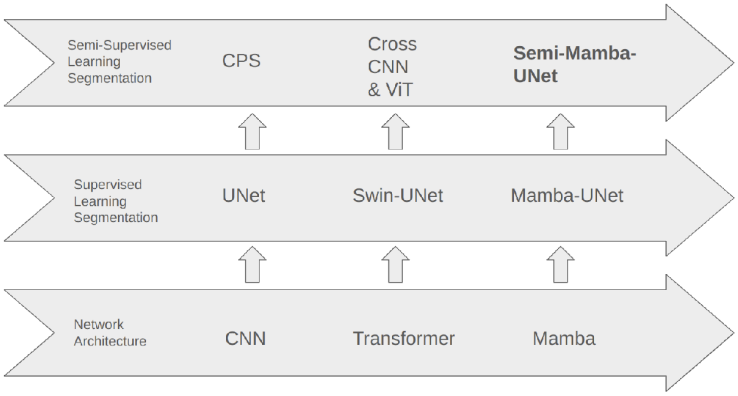

Fig. 1. The Development History of Semi-Supervised Learning, Supervised Learning for Medical Image Segmentation, and Network Architecture. Source: CNN [26], Transformer[25], Mamba[24], UNet[36], Swin-UNet[3], Mamba-UNet[52], CPS[8], Cross CNN & ViT[28], and proposed Semi-Mamba-UNet.

based on the CNN-based UNet architecture, have been extensively investigated [36, 32, 20, 48]. 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 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++ [63] introduces a nested UNet structure with deep supervision mechanisms, while Attention UNet [33] incorporates attention gates to bolster the decoders' feature learning capabilities. Moreover, Res-UNet [9] integrates residual learning [15] within its network blocks. Typically, these UNet modifications aim to leverage advanced network constructs such as densenet[19], mobilnet[17], attention mechanism[53] 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 [59, 51, 5, 22].

The recent study of multi-head self-attention from sequence-to-sequence tasks have demonstrates the effective of Transformer network architecture[41]. Image recognition is firstly been proved that Vision Transformer supervisor performance against CNN-based networks especially on the large size of dataset due to modeling long-range dependencies[10]. Several ViT-based networks hereby have been also explored in image segmentation such as SegFormer[55], Segmenter[39], and SETR[61]. In medical image segmentation domain, most of ViT-based studies have also been motivated by UNet such as TransUNet[6], SwinUNet[3], Dense SwinUNet[47], UTNet[11], Unetr[14], and nnFormer[62].

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 [27, 49, 58, 23, 50]. 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[58]. The Cross-Teaching technique extends this concept by leveraging CNN- and ViT-based UNet, enabling collaboration between the two networks through pseudo labels[28]. 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 [54].

Recent advancements have introduced the novel Mamba architecture, with strength in capturing global contextual information with efficient computational cost, conceptualized by State Space Model (SSM) [44, 13, 12]. This architecture has been explored in a variety of computer vision tasks, such as Vision Mamba [64], UMamba [31], Segmamba [56], MambaUNet [52], and VM-UNet [37]. 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, 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 medical 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 threefold:

1. Integration of a Mamba-based segmentation network with SSL for medical image segmentation. For fair evaluation, comparisons are drawn against CNN-based UNet[36] and ViT-based SwinUNet[3] across various SSL frameworks.
2. A novel pixel-level contrastive learning strategy is introduced, incorporating projector pair to maximize feature learning capabilities using provided both labeled and unlabeled data.
3. A pixel-level cross-supervised learning consists of two separate Mamba-based UNet. The network trained with the help of the other network via pseudo labeling, thereby extending the utility of unlabeled data in model training.

2 Methodology

The framework of Semi-Mamba-UNet is illustrated in Figure 2. As shown in Figure, $(\boldsymbol{X}_{\mathrm{l}}, \boldsymbol{Y}_{\mathrm{l}}) \in \mathbf{L}$ 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

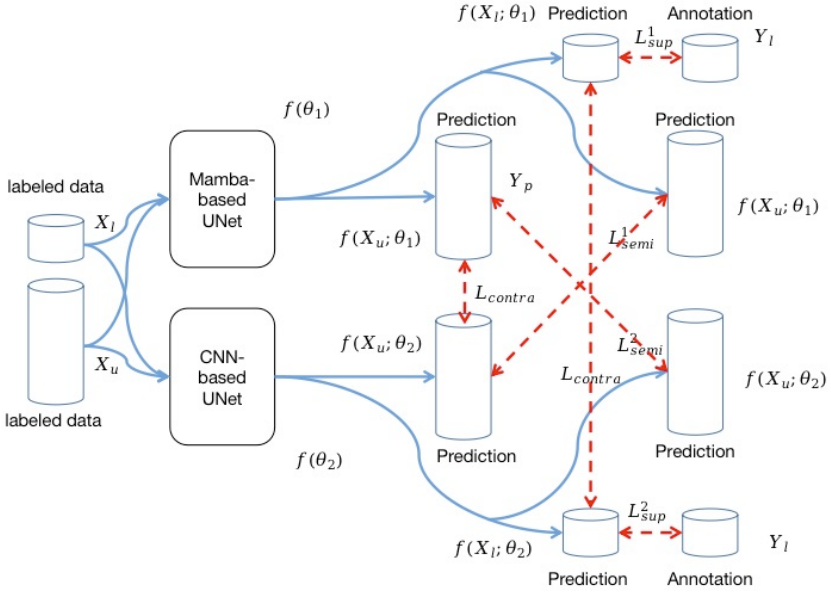

Fig. 2. Semi-Mamba-UNet: The Framework of Contrastive Cross-Supervised Visual Mamba-based UNet for Semi-Supervised Medical Image Segmentation.

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 $f(\theta_1)$ and $f(\theta_2)$. 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 $\mathcal{L}_{\mathrm{sup}}$, contrastive loss $\mathcal{L}_{\mathrm{contra}}$, and semi-supervised loss $\mathcal{L}_{\mathrm{semi}}$. The final 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 θ_1, θ_2 thus minimizing the total loss $\mathcal{L}_{total}$, illustrated as:

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

All mathematical symbols are denoted in the Figure 2 accordingly, and $\mathcal{L}$ is highlighted as red dash line, where $\mathcal{L}_{\mathrm{sup}}$ is the supervision loss for $f(\theta_1)$ and $f(\theta_2)$ based on labeled training set. The $\mathcal{L}_{\mathrm{sup}}$ is with the combination of Dice-Coefficient-based and Cross-Entropy-based loss as,

$$\mathcal{L}^1_{\mathrm{sup}} = \mathrm{CE}\big(\mathrm{argmax}(f(\boldsymbol{X}_u;\theta_1), \boldsymbol{Y}_{\mathrm{l}})\big) + \mathrm{Dice}\big(\mathrm{argmax}(f(\boldsymbol{X}_u;\theta_1), \boldsymbol{Y}_{\mathrm{l}})\big) \tag{2}$$

$$\mathcal{L}^2_{\mathrm{sup}} = \mathrm{CE}\big(\mathrm{argmax}(f(\boldsymbol{X}_u;\theta_2), \boldsymbol{Y}_{\mathrm{l}})\big) + \mathrm{Dice}\big(\mathrm{argmax}(f(\boldsymbol{X}_u;\theta_2), \boldsymbol{Y}_{\mathrm{l}})\big) \tag{3}$$

$\mathcal{L}_{\text{semi}}$ is the semi-supervision loss for $f(\theta_1)$ and $f(\theta_2)$ based on unlabeled training set. A prediction of a network is considered as the pseudo label to extend $\boldsymbol{X}_{\text{u}}$ to retrain the other networks. $\mathcal{L}_{\text{contra}}$ is the contrastive learning loss, and we proposes a projector pair to extract features between the prediction of two networks. The details of Mamba-UNet, cross-supervised learning, and contrastive learning 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 tailored for medical image analysis.

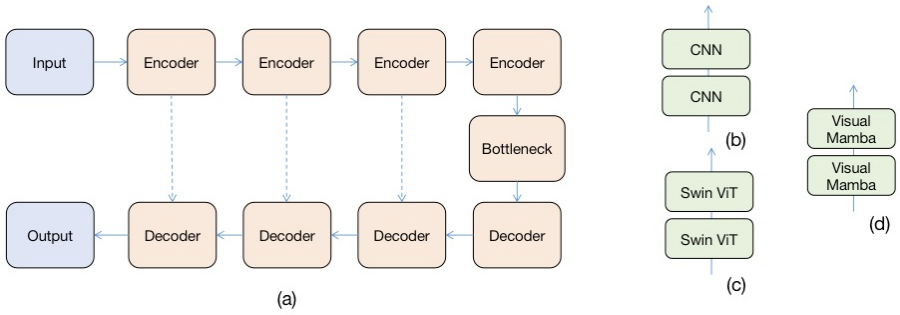

Fig. 3. The Segmentation Backbone Network in This Study. (a) Encoder-Decoder Style Segmentation Network. (b) The 2-Layer CNN-based Network Block of UNet. (c) The 2-Layer Swin ViT-based Network Block of Swin-UNet. (d) The 2-Layer Visual Mamba-based Network Block of Mamba-UNet.

To ensure a fair comparison, the proposed utilization of Mamba-UNet is developed against the original UNet [36] and the Swin UNet [3]. 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[36], the Swin-UNet utilizes 2 Swin Transformer blocks[3], and the Mamba-UNet integrates 2 Visual Mamba blocks[37, 52]. 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. 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 [8], 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 [29] 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,

$$\mathcal{L}^1_{\text{semi}} = \text{CE}\big(\text{argmax}(f(\boldsymbol{X}_u;\theta_1), f(\boldsymbol{X}_u;\theta_2))\big) + \text{Dice}\big(\text{argmax}(f(\boldsymbol{X}_u;\theta_1), f(\boldsymbol{X}_u;\theta_2))\big) \tag{4}$$

$$\mathcal{L}^2_{\text{semi}} = \text{CE}\big(\text{argmax}(f(\boldsymbol{X}_u;\theta_2), f(\boldsymbol{X}_u;\theta_1))\big) + \text{Dice}\big(\text{argmax}(f(\boldsymbol{X}_u;\theta_2), f(\boldsymbol{X}_u;\theta_1))\big) \tag{5}$$

The selection of network is based on multi-time experiments, and detailed in 'Ablation Study'.

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 [34]. The key idea of contrastive learning is the construction of an embedding space optimized to minimize the distance between analogous images while maximizing the separation between disparate ones. This technique has demonstrated remarkable efficacy across a spectrum of applications [7, 16, 21].

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 [4, 18, 57, 46]. In our proposed framework, we utilize a projector pair is appended to dual Mamba-UNet. 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 [30], formalized as,

$$\mathcal{L}_{\text{contra}} = \frac{\sum \|(G(F_\theta(z_R)), G(F_\theta(z_L)))\|_2^2 + \varepsilon}{N + \varepsilon} \tag{6}$$

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. In view of the difference in magnitude between the optimization functions, the Laplace smoothing factor is added, which is able to speed up the convergence. Here, ε is set to 10 in the experiment.

3 Experiments

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 [1]. This dataset encompasses imaging data from 100 patients, providing a comprehensive basis for evaluation. To comply with the input requirements of SwinUNet [3], 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.

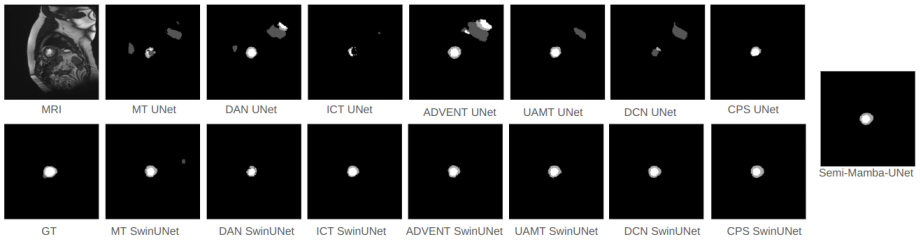

Fig. 4. The Example Segmentation Results when 5% of Data are Assumed as Labeled Data.

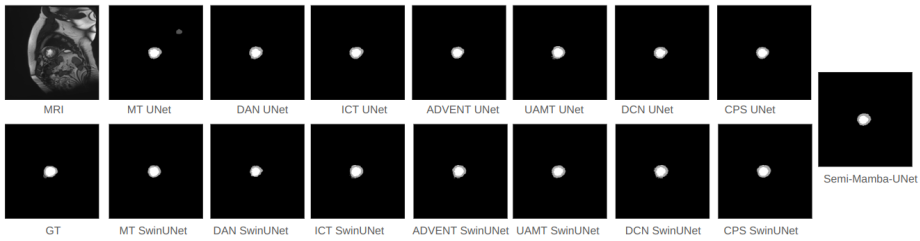

Fig. 5. The Example Segmentation Results when 10% of Data are Assumed as Labeled Data.

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 [2] 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 [36] and the Swin ViT-based SwinUNet [3] as segmentation backbone networks for different SSL frameworks. This selection was motivated by the architectural similarities these 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) [40], Deep Adversarial Network (DAN) [60], Interpolation Consistency Training (ICT) [42], Adversarial Entropy Minimization (ADVENT) [43], Uncertainty Aware Mean Teacher (UAMT) [58], and Deep Co-Training (DCN) [35]. Both SwinUNet [3] and UNet [36] were employed as the segmentation backbone networks across these frameworks.

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

Framework+Network	Dice↑	Acc↑	Pre↑	Sen↑	Spe↑	HD↓	ASD↓
MT[40] + SwinUNet	0.7506	0.9910	0.7918	0.7178	0.9394	10.4621	3.5301
DAN[60] + SwinUNet	0.7252	0.9901	0.7695	0.6903	0.9337	12.9800	4.3823
ICT[42] + SwinUNet	0.7504	0.9910	0.7923	0.7180	0.9392	9.8026	3.1055
ADVENT[43]+ SwinUNet	0.7489	0.9910	0.7964	0.7128	0.9373	11.1535	3.0907
UAMT[58] + SwinUNet[45]	0.7442	0.9909	0.7902	0.7108	0.9393	10.2955	2.8222
DCN[35] + SwinUNet	0.7603	0.9914	0.8118	0.7207	0.9376	10.1783	3.1478
CPS[8] + SwinUNet	0.7901	0.9919	0.8162	0.7730	0.9533	8.4888	2.1833
MT[40] + UNet	0.7256	0.9885	0.8217	0.6670	0.9044	24.0480	9.7662
DAN[60] + UNet	0.7657	0.9905	0.8296	0.7152	0.9199	21.1226	7.3434
ICT[42] + UNet	0.7490	0.9906	0.8827	0.6633	0.9013	11.2109	4.5181
ADVENT[43]+ UNet	0.6656	0.9833	0.6487	0.6900	0.9190	42.8011	16.6207
UAMT[58] + UNet	0.7472	0.9901	0.8164	0.6943	0.9168	21.7492	7.7489
DCN[35] + UNet	0.7312	0.9894	0.8316	0.6626	0.9022	24.6607	10.1996
CPS[8] + 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, where higher values indicate better performance ↑, include Dice Coefficient(Dice), Accuracy(Acc), Precision(Pre), Sensitivity(Sen), and Specificity(Spe). Additionally, difference measures, where lower values are preferable↓, consist of the 95% Hausdorff Distance (HD) and Average Surface

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[40] + SwinUNet	0.8678	0.9949	0.8700	0.8670	0.9745	7.3576	2.1834
DAN[60] + SwinUNet	0.8288	0.9936	0.8261	0.8375	0.9721	9.9132	2.7309
ICT[42] + SwinUNet	0.8621	0.9947	0.8624	0.8632	0.9746	8.7211	2.5562
ADVENT[43]+ SwinUNet	0.8669	0.9949	0.8688	0.8660	0.9743	7.1383	2.2608
UAMT[58] + SwinUNet[45]	0.8701	0.9950	0.8721	0.8697	0.9754	6.7226	2.0975
DCN[35] + SwinUNet	0.8608	0.9946	0.8511	0.8724	0.9777	8.8474	2.6705
CPS[8] + SwinUNet	0.8933	0.9957	0.8846	0.9032	0.9821	5.5661	1.6418
MT[40] + UNet	0.8781	0.9949	0.8836	0.8735	0.9690	10.9691	3.3246
DAN[60] + UNet	0.8766	0.9948	0.8814	0.8727	0.9700	8.6977	2.4750
ICT[42] + UNet	0.8879	0.9953	0.8996	0.8779	0.9696	6.7011	1.9696
ADVENT[43]+ UNet	0.8777	0.9949	0.8877	0.8703	0.9674	11.0979	2.9367
UAMT[58] + UNet	0.8798	0.9949	0.8778	0.8821	0.9726	10.2134	3.1926
DCN[35] + UNet	0.8831	0.9952	0.8897	0.8785	0.9706	8.6978	2.7026
CPS[8] + 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**

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**

Distance (ASD). Given that the dataset involves 4-class segmentation tasks, the mean values of 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>.

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 Mamba-based 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 cross-supervision with self-supervised 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.

References

1. Bernard, O., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
2. Bottou, L.: Stochastic gradient learning in neural networks. In: Proceedings of Neuro-Nîmes 91. EC2, Nimes, France (1991)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
4. Chaitanya, K., et al.: Contrastive learning of global and local features for medical image segmentation with limited annotations. NIPS (2020)
5. Chaurasia, A., Culurciello, E.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP). pp. 1–4. IEEE (2017)
6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
7. Chen, T., et al.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
8. Chen, X., et al.: Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR (2021)
9. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing **162**, 94–114 (2020)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

11. Gao, Y., Zhou, M., Metaxas, D.N.: 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. pp. 61–71. Springer (2021)
12. Gu, A.: Modeling Sequences with Structured State Spaces. Ph.D. thesis, Stanford University (2023)
13. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
14. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. He, K., et al.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
17. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
18. Hu, X., et al.: Semi-supervised contrastive learning for label-efficient medical image segmentation. In: MICCAI. Springer (2021)
19. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
20. Ibtehaz, N., Rahman, M.S.: Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. Neural networks **121**, 74–87 (2020)
21. Kang, G., et al.: Contrastive adaptation network for unsupervised domain adaptation. In: CVPR. pp. 4893–4902 (2019)
22. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)
23. Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A.: Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems **32**(2), 523–534 (2020)
24. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)
25. Liu, Z., Lin, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
27. Luo, X., Hu, M., Liao, W., Zhai, S., Song, T., Wang, G., Zhang, S.: 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. pp. 528–538. Springer (2022)
28. Luo, X., et al.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. arXiv preprint arXiv:2112.04894 (2021)
29. Luo, X., et al.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: MIDL (2022)

30. Ma, C.: Ssl-unet: A self-supervised learning strategy base on u-net for retinal vessel segmentation. Authorea Preprints (2023)
31. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
32. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
33. Oktay, O., et al.: Attention U-Net: Learning where to look for the pancreas. Int Conf Medical Imaging with Deep Learning (2018)
34. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
35. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: Proceedings of the European Conference on Computer Vision. pp. 135–152 (2018)
36. Ronneberger, O., et al.: U-Net: Convolutional networks for biomedical image segmentation. In: Int Conf Med Im Comp & Comp-Assisted Intervention. pp. 234–241. Springer (2015)
37. Ruan, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)
38. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
39. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262–7272 (2021)
40. Tarvainen, A., Valpola, H.: 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. pp. 1195–1204 (2017)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
42. Verma, V., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. In: International Joint Conference on Artificial Intelligence. pp. 3635–3641 (2019)
43. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2517–2526 (2019)
44. Wang, J., Zhu, W., Wang, P., Yu, X., Liu, L., Omar, M., Hamid, R.: Selective structured state-spaces for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6387–6397 (2023)
45. Wang, Z.: An uncertainty-aware transformer for mri cardiac semantic segmentation via mean teachers. Annual Conference on Medical Image Understanding and Analysis (2022)
46. Wang, Z., Ma, C.: 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. pp. 870–879 (2023)

47. Wang, Z., Su, M., Zheng, J.Q., Liu, Y.: Densely connected swin-unet for multiscale information aggregation in medical image segmentation. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 940–944. IEEE (2023)
48. Wang, Z., Voiculescu, I.: 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) (2021)
49. Wang, Z., Voiculescu, I.: 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. pp. 181–190. Springer (2023)
50. Wang, Z., Voiculescu, I.: Weakly supervised medical image segmentation through dense combinations of dense pseudo-labels. In: MICCAI Workshop on Data Engineering in Medical Imaging. pp. 1–10. Springer (2023)
51. Wang, Z., et al.: 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)
52. Wang, Z., et al.: Mamba-unet: Unet-like pure visual mamba for medical image segmentation. arXiv preprint arXiv:2402.05079 (2024)
53. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
54. Xia, Y., Liu, F., Yang, D., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H.: 3d semi-supervised learning with uncertainty-aware multi-view co-training. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3646–3655 (2020)
55. Xie, E., et al.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
56. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)
57. You, C., et al.: Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE TMI (2022)
58. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–613. Springer (2019)
59. Zhang, Y., Yuan, L., Wang, Y., Zhang, J.: Sau-net: efficient 3d spine mri segmentation using inter-slice attention. In: Medical Imaging With Deep Learning. pp. 903–913. PMLR (2020)
60. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: International conference on medical image computing and computer-assisted intervention. pp. 408–416. Springer (2017)
61. Zheng, S., 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. pp. 6881–6890 (2021)
62. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE Transactions on Image Processing (2023)

63. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11. Springer (2018)
64. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)