Weak-Mamba-UNet: Visual Mamba Makes CNN and ViT Work Better for Scribble-based Medical Image Segmentation

Ziyang Wang* *Senior Member, IEEE*, Tianli Tao *Student Member, IEEE*, Yiyuan Ge *Student Member, IEEE*, Zhihao Chen *Student Member, IEEE*, Tianxiang Chen *Student Member, IEEE*, Zi Ye *Member, IEEE*, Yongxiang Lei *Member, IEEE*

***Abstract*— Medical image segmentation is increasingly reliant on deep learning techniques, yet the promising performance often come with high annotation costs. This paper introduces Weak-Mamba-UNet, an innovative weakly-supervised learning (WSL) framework that leverages the capabilities of Convolutional Neural Network (CNN), Vision Transformer (ViT), and the cutting-edge Visual Mamba (VMamba) architecture for medical image segmentation, especially when dealing with scribble-based annotations. The proposed WSL strategy incorporates three distinct architecture but same symmetrical encoder-decoder networks: a CNN-based U-Net for detailed local feature extraction, a Swin Transformer-based Swin-UNet for comprehensive global context understanding, and a VMamba-based Mamba-UNet for efficient long-range dependency modeling. The key concept of this framework is a collaborative and cross-supervisory mechanism that employs pseudo labels to facilitate iterative learning and refinement across the networks. The effectiveness of Weak-Mamba-UNet is validated on two publicly available datasets with processed scribble annotations, where it surpasses the performance of a similar WSL framework utilizing only U-Net or Swin-UNet, as well as other baseline methods. This paper highlights the potential of Mamba for medical image segmentation in scenarios with sparse or imprecise annotations. The source code, dataset, and all baseline methods are made publicly accessible https://github.com/ziyangwang007/Mamba-UNet.**

***Index Terms*— Visual Mamba, U-Net, Weak-Supervised Learning, Scribble, Medical Image Segmentation**

Ziyang Wang is with School of Computer Science and Digital Technologies, Aston University, Biringham, UK. Ziyang Wang is the corresponding author. (Email: z.wang47@aston.ac.uk)

Tianli Tao is with School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK. (Email: tianli.tao@kcl.ac.uk)

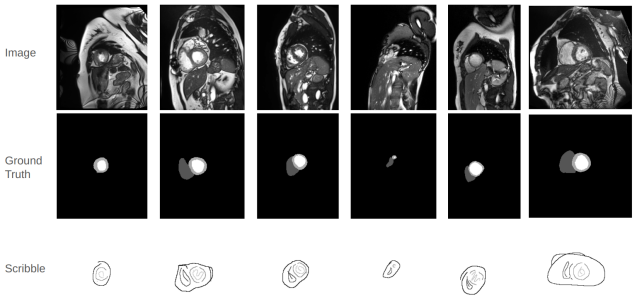
Yiyuan Ge is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China. (Email: 202510182871@mail.scut.edu.cn)

Zhihao Chen is with the School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing, China. (Email: zhihaochen666@bupt.edu.cn)

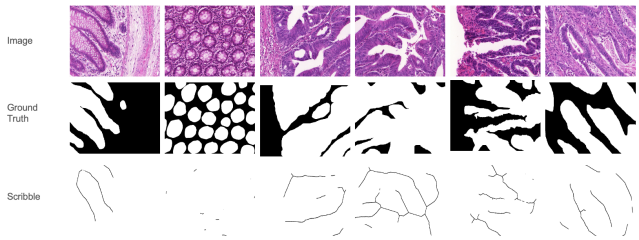
Tianxiang Chen is with School of Cyber Science and Technology, University of Science and Technology of China, Hefei, China. (Email: txchen@mail.ustc.edu.cn)

Zi Ye is with Trinity Institute of Neurosciences, Trinity College Dublin, The University of Dublin, Dublin, Ireland. (Email: yez3@tcd.ie)

Yongxiang Lei is with Intelligent Control & Smart Energy (ICSE) Research Group, School of Engineering, University of Warwick, Coventry, UK. (Email: yongxiang.lei@warwick.ac.uk)



(a) MRI cardiac scans with corresponding segmentation ground truth and scribble-based annotations.



(b) Colon histology images with corresponding gland segmentation ground truth and scribble-based annotations.

Fig. 1. Example images of different medical modalities: (a) MRI cardiac scans and (b) colon histology images, each with segmentation ground truth and scribble-based annotations.

I. Introduction

Medical image segmentation is important for medical image analysis and effective treatment planning for healthcare purpose, with deep learning-based networks i.e. U-Net [38]. The U-Net known for its symmetrical U-shape encoder-decoder architecture and integral skip connections, has been the foundational segmentation backbone network. These skip connections effectively preserve essential spatial information, merging features across the encoder and decoder layers to enhance the network's performance. The encoder reduces the input to extract high-level features, which the decoder then uses to reconstruct the image, thereby improving segmentation performance. Advancements in U-Net have led to various enhanced networks designed to tackle the segmentation of

complex anatomical structures in Ultrasound, CT and MRI scans [35], [51], [66], [73], [75].

Recent advancements have introduced innovative architectures such as the Transformer and Mamba, both of which excel in capturing global contextual information [13], [44]. The Transformer achieves this through a multi-head self-attention mechanism, while Mamba is noted for its computational efficiency, grounded in the State Space Model (SSM) [12], [13], [49]. These architectures have been applied to a range of computer vision tasks, leading to developments like the Vision Transformer [10], Swin Transformer [26], nnFormer [74], ScribFormer [22], and UNetr [15] for Transformer, and Vision Mamba [76], UMamba [33], Segmamba [62], MambaUNet [52], VM-UNet [39], and Semi-MambaUNet [54] for Mamba-based networks.

The effectiveness of deep learning methods often hinges on the availability of large, accurately labeled datasets, which can be challenging to acquire in the medical image analysis domain. To address the high costs and time associated with obtaining detailed annotations like pixel-level segmentation masks, research has shifted towards Semi-Supervised Learning (SSL) [5], [17], [29], [58] and Weakly-Supervised Learning (WSL) [24], [31], [36], [59], [60]. SSL focuses on training networks with a small set of pixel-level labeled data, whereas WSL employs simpler forms of annotations such as bounding boxes, checkmarks, and points to provide a feasible approach for training segmentation networks under limited-signal supervision. Among these, scribble-based annotation is particularly noted for its efficiency and convenience for experts, streamlining the annotation process without significantly compromising the quality of supervision. Examples of MRI scans, conventional dense annotations, and scribble-based annotations are illustrated in Figure 1.

Following the recent success of the Transformer and Mamba architectures in computer vision tasks, and concern with limited annotated data, this paper introduces Weak-Mamba-UNet. The proposed WSL framework integrates Convolution, Transformer, and Mamba architectures within a multi-view cross-supervised learning scheme tailored for scribble-based supervised medical image segmentation. To the best of our knowledge, this is the first effort to leverage the Mamba architecture for medical image segmentation with scribble annotations. The contributions of Weak-Mamba-UNet are threefold:

1) We propose Weak-Mamba-UNet, a tri-view cross-supervision framework that jointly trains three heterogeneous encoder–decoder backbones—U-Net (CNN), Swin-UNet (ViT), and Mamba-UNet (state-space)—to exploit complementary inductive biases under scribble supervision.
2) We introduce a stochastic soft pseudo-label mixing mechanism: at each iteration, dense targets are created by Dirichlet-sampled mixing of peer probability maps and optimized with partial cross-entropy (on scribbles) + soft Dice (on the dense soft targets). This converts sparse scribbles into strong supervision without class embeddings, CRFs, or additional priors.
3) We show that cross-supervising across architectures—rather than duplicating a single backbone—yields consistent gains on ACDC cardiac MRI (scribbles derived from dense labels), with improvements over strong scribble-WSL baselines and over tri-homogeneous variants (3×U-Net and 3×Mamba-UNet).

II. Related Work

A. Medical Image Segmentation

Medical image segmentation is a core task in medical image analysis that aims to partition images into anatomical or pathological regions of interest [1], [11], [16], [38], [51], [56], [66], [72]. Early advances were driven by fully convolutional architectures such as FCN [27] and, especially, U-Net [38], whose symmetric encoder–decoder with skip connections preserves spatial detail critical for clinical structures. Numerous variants further improve feature propagation, volumetric modeling, and attention to relevant regions, e.g., U-Net++ [75], V-Net [35], and attention-enhanced designs [37], [40]. These ideas have been explored broadly across modalities including ultrasound, CT, and MRI [35], [51], [66], [73], [75].

Transformer-based models have recently been adapted to segmentation to better capture long-range dependencies and global context [10], [26]. Integrations of Transformers with U-shaped decoders—such as Mixed Transformer UNet [47], Swin-UNet [3], and MCV-UNet [65]—combine convolutional locality with attention-based global reasoning, and continue to improve dense prediction quality in medical images [3], [55].

B. Weakly-Supervised Medical Image Segmentation

Obtaining dense pixel-level annotations is costly and time-consuming in clinical practice; weakly supervised learning (WSL) therefore leverages cheaper supervision such as boxes, scribbles, image-level tags, text, and points [4], [6], [8], [14], [19], [20], [24], [28], [31], [36], [50], [59], [60], [68]. Among these, scribbles are attractive because they provide sparse yet informative cues with minimal effort. Effective use of scribbles typically requires (i) turning sparse cues into dense supervision via pseudo labels [31], [43], [60] and (ii) regularizing predictions to align with annotated pixels while maintaining spatial coherence, e.g., with partial cross-entropy [41] and CRF-based smoothing [36]. In addition, multi-view or multi-scale consistency encourages robustness by requiring predictions to agree across different perturbations or perspectives [9], [31], [53], [57]. Recent robustness benchmarks such as RAOS (Rethinking Abdominal Organ Segmentation) emphasize that clinically realistic 'hard cases' can substantially degrade performance, underscoring the need for dedicated robustness evaluation before deployment [32].

Relation to cross-supervision.: Cross pseudo supervision and cross-teaching enforce agreement between parallel learners trained under different perturbations or decoders [5], [30]. ScribbleVC [21] builds a hybrid CNN–Transformer with two decoders and a vision–class embedding module, while dual-branch dynamic mixed pseudo-labeling methods operate within a single architecture family [31]. Our framework is related in spirit but differs in *scope*: we consider heterogeneous CNN/ViT/Mamba backbones and use stochastic mixing of

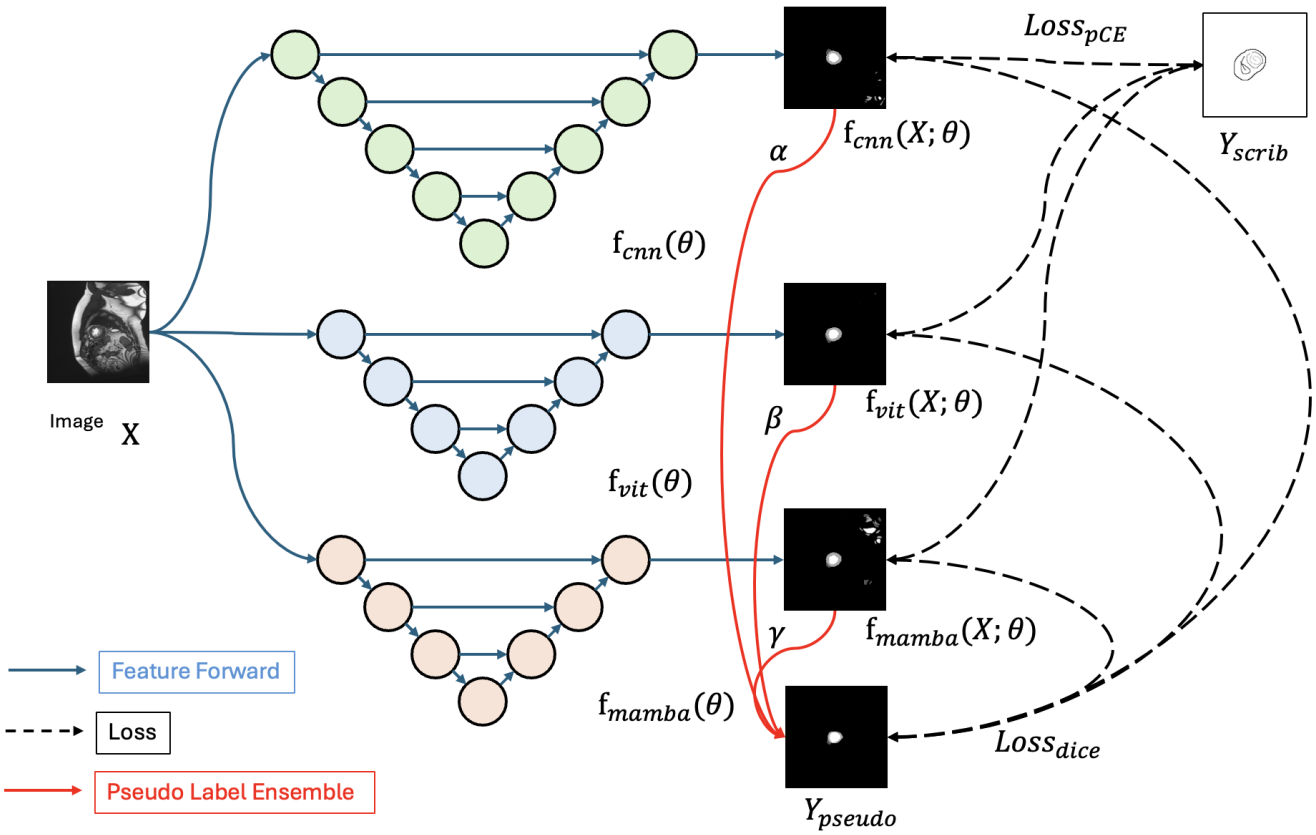


Fig. 2. Weak-Mamba-UNet framework overview (see Methodology for details).

peer predictions to obtain dense signals. Detailed algorithmic choices are described in Methodology.

C. Visual Mamba for Medical Image Segmentation

State space models (SSMs) provide linear-time sequence operators that have recently been adapted to vision as Mamba/VMamba [13], [25], [76], offering efficient long-range dependency modeling [63], [70]. Vision Mamba has been instantiated in diverse architectures, including U-Mamba [33], Segmamba [62], Polyp-Mamba [64], H-vmunet [61], Swin-UMamba [23], LKM-UNet [48], and Mamba-UNet [52], showing strong performance in classification, detection, and segmentation [25], [76]. Given these advantages, applying Mamba to medical segmentation is promising and has begun to show benefits across tasks and modalities [54], [67], [69].

III. Methodology

The framework of Weak-Mamba-UNet is illustrated in Figure 2. In this study, the pair $(\mathbf{X}, \mathbf{Y}_{\text{scrib}})$ represents the scribble-based labeled training dataset, whereas the pair $(\mathbf{X}_t, \mathbf{Y}_t)$ denotes the dense labeled testing dataset. Here, $\mathbf{X} \in \mathbb{R}^{h \times w}$ corresponds to a 2D grayscale image of height h and width w. The scribble annotations $\mathbf{Y}_{\text{scrib}} \in \{0, 1, 2, 3, \text{None}\}$ indicate the regions corresponding to the right ventricle (RVC), left ventricle (LVC), myocardium (MYO), background, and unlabeled pixels, respectively.

Three segmentation networks are denoted as $f_{\text{cnn}}(\mathbf{X};\theta)$, $f_{\text{vit}}(\mathbf{X};\theta)$, and $f_{\text{mamba}}(\mathbf{X};\theta)$, and are highlighted in green, blue, and yellow in Figure 2 as the complete framework, and Figure 4 as each of network block, respectively. The prediction of a segmentation network for an input $\mathbf{X}$ is denoted as $\mathbf{Y}_p = f(\mathbf{X};\theta)$, where θ represents the network parameters. The predictions from the three networks can be combined to form a dense pseudo label $\mathbf{Y}_{\text{pseudo}}$.

The overall loss comprises the scribble-based partial cross-entropy loss $\mathcal{L}_{\text{pce}}$ and the dense-signal pseudo label dice-coefficient loss $\mathcal{L}_{\text{dice}}$. The total training objective aims to minimize the combined loss $\mathcal{L}_{\text{total}}$, which is formulated as:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{3} (\mathcal{L}_{\text{pce}}^{i} + \mathcal{L}_{\text{dice}}^{i}), \tag{1}$$

where i indicates each of three networks. All mathematical symbols are defined in Figure 2. The final evaluation assesses the agreement between the predicted labels $\mathbf{Y}_p$ and the true dense labels $\mathbf{Y}_t$ on the test set.

A. Preliminaries

State Space Models (SSMs) provide a principled framework for sequence modeling by drawing inspiration from continuous-time linear time-invariant (LTI) systems. These systems map a 1D input sequence $x(t) \in \mathbb{R}$ to an output $y(t) \in \mathbb{R}$ through an implicit latent state $h(t) \in \mathbb{R}^N$, governed by linear ordinary differential equations (ODEs):

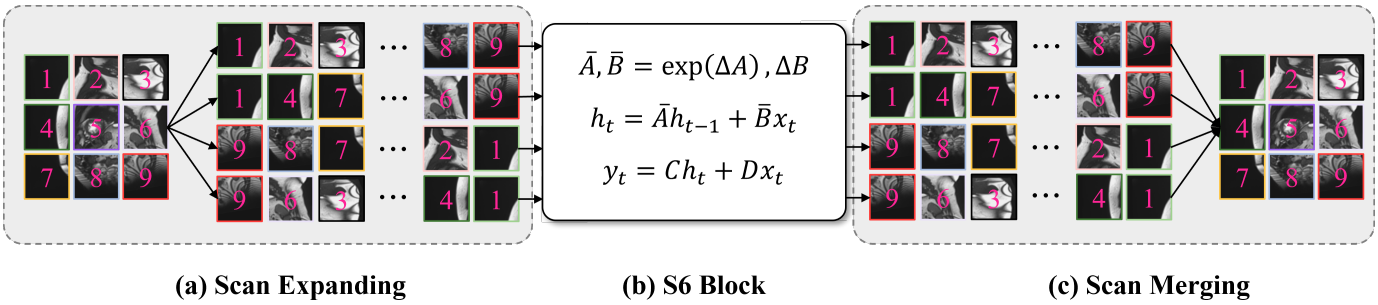


Fig. 3. The Illustration of the 2D Selective Scan (SS2D) on a cardiac image. We first apply a scan-expanding process to generate four feature sequences. Each sequence is then processed through the S6 block, and the resulting outputs are merged (scan merging) to form the final 2D feature map.

$$\begin{cases} h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) = \mathbf{C}h(t) + \mathbf{D}x(t), \end{cases} \tag{2}$$

where N is the state size, $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the state transition matrix, $\mathbf{B} \in \mathbb{R}^{N}$ and $\mathbf{C} \in \mathbb{R}^{N}$ are projection parameters, $\mathbf{D} \in \mathbb{R}$ is the skip connection parameter.

State Space Models (SSMs), while theoretically grounded in continuous-time dynamics, face significant implementation challenges in deep learning due to their inherent *continuous-time nature*. To model these dynamics in discrete time, a zero-order hold is applied, resulting in the following system, where $\Delta \in \mathbb{R}^{D}$ represents the timescale parameter:

$$\begin{aligned} h_t &= \overline{A}h_{t-1} + \overline{B}x_t, \\ y_t &= Ch_t + \overline{D}x_t, \\ \overline{A} &= e^{\Delta A}, \\ \overline{B} &= (e^{\Delta A} - I)A^{-1}B \end{aligned} \tag{3}$$

In this formulation, $\overline{B}$ is approximated using a first-order Taylor expansion, which simplifies the calculation as $\overline{B} \approx \Delta B$. This approximation maintains computational efficiency while preserving the ability to capture complex temporal and spatial dependencies in the data.

However, directly applying Mamba to visual tasks faces inherent limitations due to the dimensional incompatibility between language sequences and 2D visual data. While 1D modeling efficiently captures long-range dependencies in text, it fails to preserve spatial relationships critical for images, resulting in limited receptive fields and neglected cross-region correlations. To address this, the 2D Selective Scan (SS2D) framework [25] extends Mamba through spatial adaptation. As shown in Figure 3, the method unfolds input images into scanning sequences along four diagonal directions. This directional diversity ensures each feature element dynamically integrates cross-orientation contextual information, constructing global receptive fields while preserving linear computational scaling.

B. Visual Mamba-Based U-Net

The VMamba-Based block along with other version of network block, as depicted in the Figure 4, represents a novel adaptation to the conventional U-Net encoder-decoder style segmentation network in Figure 2, incorporating various types of network blocks specifically designed for medical image analysis. Each of these architectures adheres to the U-shaped encoder-decoder configuration, with distinct block compositions that influence the network's capability to capture intricate medical image features. In particular, the original U-Net employs a two-layer convolutional neural network (CNN) with 3×3 convolutions, Swin-UNet utilizes two Swin Transformer blocks, and Mamba-UNet integrates two Visual Mamba blocks which is an advanced module introduced in our previous work [52].

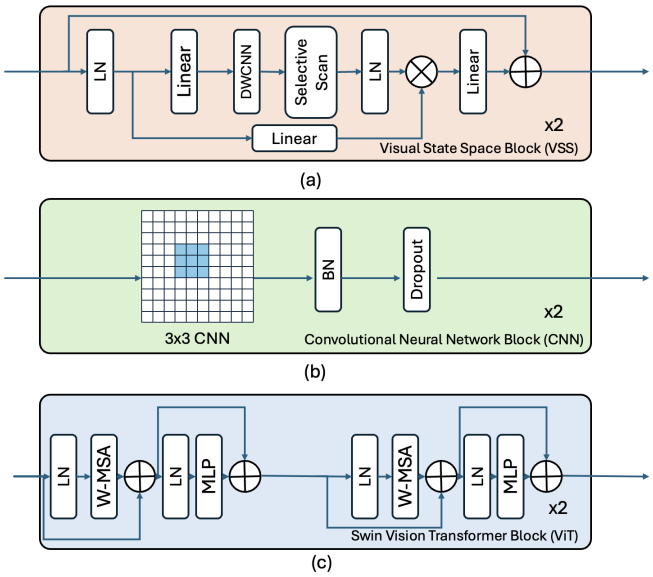


Fig. 4. The Illustration of Each Network Block Utilized for the Framework of Weak-Mamba-UNet.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, a 1×1 projection first maps channels to C' (keeping $H\times W$ unchanged). The SS2D module then *unfolds* $\mathbf{X}$ into four directional scan sequences corresponding to the diagonal orientations as illustrated in Fig. 3. Each directional sequence is processed by the *same* S6 (Mamba) operator, producing four outputs that are *folded* back to $B\times C'\times H\times W$ and *merged* by element-wise averaging. A final 1×1 projection restores the channel dimension to C, followed by a residual addition with the block input. Skip connections and the down/up-sampling operations

of U-Net remain unchanged. This drop-in design preserves the U-shaped topology while enabling linear-time, directional long-range aggregation inside each Mamba block. We share the S6 (Mamba) parameters across the four directions to keep the block lightweight and rotationally consistent; fusion is a simple average (no extra gates), followed by a 1×1 projection and residual addition.

C. Scribble-Supervised Learning

To address the challenges posed by sparse-signal scribble supervision, we utilize a modified CrossEntropy (CE) function that concentrates solely on the annotated pixels while ignoring the unlabeled ones. This approach leads to a form of partially supervised segmentation loss. Specifically, we introduce the Partial Cross-Entropy (pCE) [41], which leverages only the scribble annotations during the training of the networks, denoted as $\mathcal{L}_{\text{pce}}$. This is expressed in Equation 4 as follows:

$$\mathcal{L}_{\text{pce}} = -\sum_{i\in\Omega_L}\sum_{k} y_{\text{s}}[i,k]\log(y_{\text{p}}[i,k]), \tag{4}$$

where i denotes the index of a given pixel, and Ω_L represents the set of pixels annotated with scribbles. The variable k indicates the class index(4 in this study), and $y_{\text{s}}[i,k]$ and $y_{\text{p}}[i,k]$ denote the ground truth and predicted probability of a network, respectively, of the i-th pixel belonging to the k-th class. The $\mathcal{L}_{\text{pce}}$ is utilized for all three networks $f_{\text{cnn}}(\mathbf{X};\theta)$, $f_{\text{vit}}(\mathbf{X};\theta)$, and $f_{\text{mamba}}(\mathbf{X};\theta)$, and denoted as $\mathcal{L}^{i}_{\text{pce}}$ where $i \in [1,2,3]$,

D. Multi-View Cross-Supervised Learning

Inspired by Cross Pseudo Supervision (CPS) [5], Cross Teaching [30], and Multi-view Learning [57], which are designed to facilitate consistency regularization under different network perturbations, our proposed multi-view cross-supervised learning framework integrates Mamba-UNet [52] with the original U-Net [38] and Swin-UNet [3]. Each network follows a U-shaped encoder-decoder architecture. Specifically, U-Net employs a 2-layer CNN with 3×3 kernels [38] and performs 4 levels of downsampling and upsampling. Swin-UNet utilizes 2 Swin Transformer blocks [3], and Mamba-UNet incorporates 2 Visual Mamba blocks [39], [52]. Both Swin-UNet and Mamba-UNet perform 3 levels of downsampling and upsampling and are pretrained on ImageNet [7]. This setup introduces three distinct architectural perspectives, each initialized separately to ensure diversity in viewpoints. To foster mutual enhancement among the networks, a composite pseudo label $\mathbf{Y}_{\text{pseudo}}$ is formulated to convert sparse-label information into dense signal labels, as shown in the equation below:

$$\mathbf{Y}_{\text{pseudo}} = \alpha\,\mathbf{p}_{\text{cnn}} + \beta\,\mathbf{p}_{\text{vit}} + \gamma\,\mathbf{p}_{\text{mamba}}, \tag{5}$$

where $\mathbf{p}_{\cdot} = \mathrm{softmax}(f_{\cdot}(\mathbf{X}))$ are class-probability maps. This *architecturally heterogeneous* mixing (CNN+ViT+Mamba) serves as a view perturbation akin to mixup-style regularization, but acts on *model predictions* rather than input pixels.Training minimizes a sum of (i) *partial cross-entropy* on the scribbled index set Ω_L, and (ii) a *soft-Dice* loss against the soft pseudo label in Eq. 5:

$$\mathcal{L}_{\text{pce}} = -\sum_{i\in\Omega_L}\sum_{k} y_s[i,k]\log p[i,k], \tag{6}$$

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_{i,k} p[i,k]\;\tilde{y}[i,k]}{\sum_{i,k} p[i,k] + \sum_{i,k}\tilde{y}[i,k] + \epsilon}, \tag{7}$$

where p is the prediction of the supervised backbone at that step and $\tilde{y} = \mathbf{Y}_{\text{pseudo}}$. The total objective is $\mathcal{L}_{\text{total}} = \sum_j(\mathcal{L}^{(j)}_{\text{pce}} + \lambda\,\mathcal{L}^{(j)}_{\text{dice}})$, with λ controlling the contribution of the dense pseudo-label supervision. Unlike ScribbleVC, we do not use a class-embedding module; all dense signals arise from peer predictions.

Unlike dual-decoder consistency methods within a single backbone, our supervision graph spans heterogeneous CNN/ViT/Mamba learners. They are trained simultaneously on the same mini-batch; their parameters are disjoint and no gradients are shared. The soft target $\mathbf{Y}_{\text{pseudo}}$ (Eq. 5) is obtained by Dirichlet-sampled mixing of peer probability maps, which acts as an architectural view perturbation and regularizer during training. This differs from (i) test-time ensembling (which averages final predictions but does not provide dense training signals), and (ii) class-embedding or CRF-based densification (which rely on external priors/graphs). Empirically (Table IV), cross-architecture supervision outperforms tri-homogeneous variants under the same scribble WSL protocol, suggesting that the gains stem from complementary inductive biases rather than stacking capacity.

IV. Experiments

A. Datasets

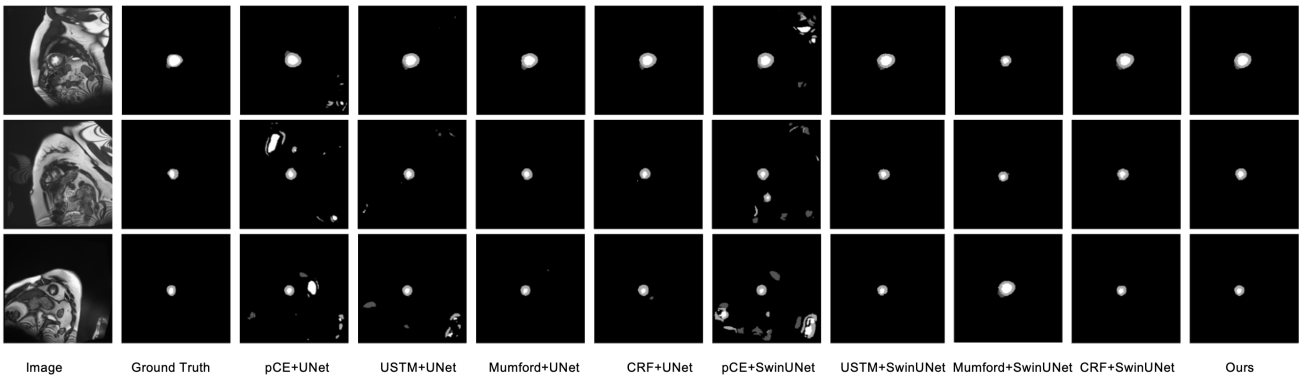
The performance of Weak-Mamba-UNet, as well as various baseline methods, were evaluated using a publicly available MRI cardiac segmentation dataset [1] and PH2 [34]. Scribble annotations were derived from the original dense annotations, in line with previous studies [43]. All images were resized to a uniform resolution of 224×224 pixels for consistency in the evaluation process.

B. Implementation Details

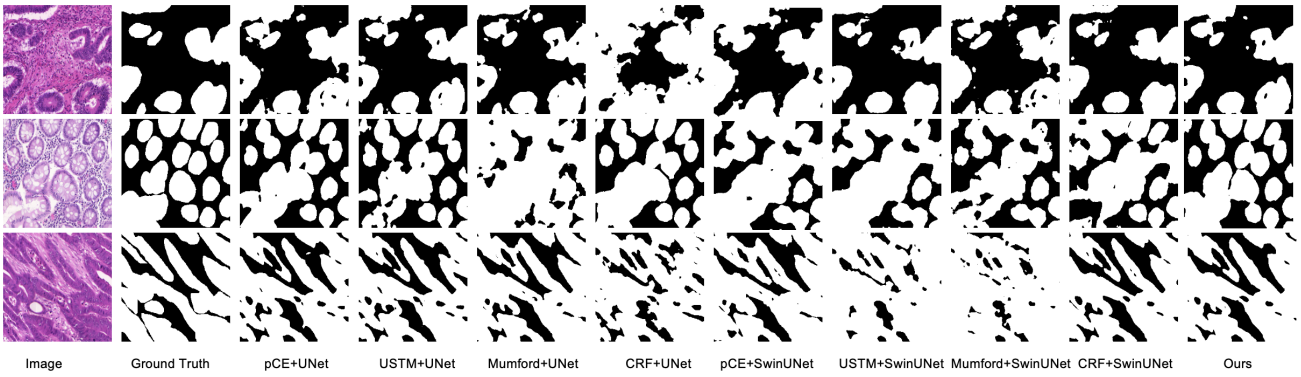
The experiments were conducted on an Ubuntu 20.04 system equipped with an Nvidia GeForce RTX 3090 GPU and an Intel Core i9-10900K CPU, using PyTorch. The entire experimental run took an average of 4 hours. We trained Weak-Mamba-UNet with all other baseline methods for 30,000 iterations with a batch size of 24. Optimization was performed using Stochastic Gradient Descent (SGD) [2], with an initial learning rate of 0.01, momentum set to 0.9, and weight decay at 0.0001. The networks were evaluated on the validation set every 200 iterations, saving the network weights only when the validation performance improved.

C. Baseline Methods

The framework of Weak-Mamba-UNet is depicted in Figure 2 with three segmentation backbone networks. To ensure



(a) Example segmentation results of all baseline methods and Weak-Mamba-UNet on the ACDC test set.



(b) Example segmentation results of all baseline methods and Weak-Mamba-UNet on the GlaS test set.

Fig. 5. Qualitative segmentation comparisons across datasets. (a) ACDC test set and (b) GlaS test set, showing the performance of baseline methods and the proposed Weak-Mamba-UNet.

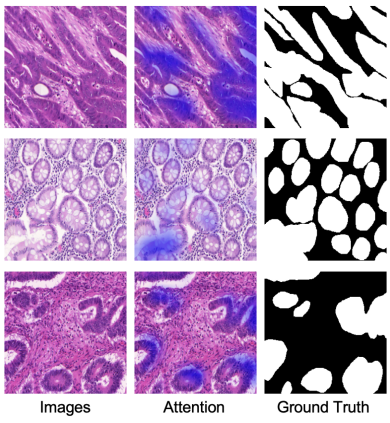


Fig. 6. The visualization of the attention heat maps (blue) generated by Weak-Mamba-UNet, along with the corresponding input images and ground truth segmentation.

equitable comparisons, we also employed the CNN-based U-Net [38] and the Swin ViT-based Swin-UNet [3] as segmentation backbone networks for different WSL frameworks. The WSL baseline frameworks evaluated includes partial Cross Entropy (pCE) [42], Uncertainty-aware Self-ensembling and Transformation-consistent Mean Teacher Model (USTM) [71], Mumford [45], ScribFormer [22], ScribbleVC [21], Gated Conditional Random Field (Gated CRF) [46]. Both Swin-UNet [3] and U-Net [38] were employed as the segmentation backbone networks across these frameworks. All baseline methods and our proposed methods are with the same hyperparameter settings.

D. Metrics

To evaluate the performance of Weak-Mamba-UNet relative to other WSL baseline methods, we employed a set of comprehensive evaluation metrics. For similarity measures, where higher values indicate better performance (↑), we included the Dice Coefficient (Dice), Accuracy (Acc), Precision (Pre), Sensitivity (Sen), and Specificity (Spe). For difference measures, where lower values are preferable (↓), we considered the 95% Hausdorff Distance (HD) and Average Surface Distance (ASD). Given the dataset's focus on 4-class segmentation tasks, we report the mean values of these metrics across all classes.

E. Computational Analysis

To further substantiate the claim of efficiency, we compare the computational cost of the three backbone networks in Table VI. All experiments were conducted on an NVIDIA RTX 3090 GPU under identical settings. Compared with the Transformer-based Swin-UNet, the Mamba-UNet achieves a **59% reduction in FLOPs**, **63% fewer parameters**, and requires **21% less GPU memory**, while also reducing both training and inference time. These results confirm that the

TABLE I
DIRECT COMPARISON OF WEAK-SUPERVISED FRAMEWORKS ON MRI CARDIAC TEST SET.

Framework+Network	Dice↑	Acc↑	Pre↑	Sen↑	Spe↑	HD_{95} ↓	ASD↓
pCE [41] + U-Net	0.7620	0.9807	0.6799	0.9174	0.9823	151.0593	54.6531
USTM [24] + U-Net	0.8592	0.9917	0.8128	0.9257	0.9888	99.8293	26.0185
Mumford [18] + U-Net	0.8993	0.9950	0.8844	0.9200	0.9874	28.0604	7.3907
Gated CRF [36] + U-Net	0.9046	0.9955	0.8890	0.9304	0.9922	7.4340	2.0753
pCE [41] + Swin-UNet	0.8935	0.9950	0.8808	0.9129	0.9884	24.4750	6.9108
USTM [24] + Swin-UNet	0.9044	0.9957	0.8952	0.9187	0.9898	6.5172	2.2319
Mumford [18] + Swin-UNet	0.9051	0.9958	0.8996	0.9157	0.9889	6.0653	1.6482
Gated CRF [36] + Swin-UNet	0.8995	0.9955	0.8920	0.9175	0.9904	6.6559	1.6222
ScribFormer [22]	0.8420						
ScribbleVC [21]	0.895						
Weak-Mamba-UNet	0.9171	0.9963	0.9095	0.9309	0.9920	3.9597	0.8810

TABLE II
DIRECT COMPARISON OF WEAK-SUPERVISED FRAMEWORKS ON GLAS HISTOLOGICAL TEST SET.

Framework+Network	Dice↑	Acc↑	Pre↑	Sen↑	Spe↑	HD_{95} ↓	ASD↓
pCE [41] + U-Net	0.8381	0.9837	0.7813	0.9038	0.9813	42.7315	10.3842
USTM [24] + U-Net	0.8791	0.9862	0.8541	0.9057	0.9849	33.2846	8.9431
Mumford [18] + U-Net	0.9031	0.9924	0.8897	0.9169	0.9872	26.9152	7.2864
Gated CRF [36] + U-Net	0.9095	0.9951	0.8965	0.9228	0.9896	21.7643	6.2049
pCE [41] + Swin-UNet	0.9023	0.9936	0.8902	0.9147	0.9880	24.5087	6.7825
USTM [24] + Swin-UNet	0.9136	0.9954	0.9063	0.9211	0.9891	20.9361	6.0137
Mumford [18] + Swin-UNet	0.9192	0.9957	0.9126	0.9258	0.9889	19.8742	5.7843
Gated CRF [36] + Swin-UNet	0.9187	0.9961	0.9072	0.9305	0.9898	18.9526	5.4628
ScribFormer [22]	0.9103						
ScribbleVC [21]	0.8942						
Weak-Mamba-UNet	0.9360	0.9964	0.9273	0.9449	0.9907	17.4368	4.9156

TABLE III
ABLATION STUDIES ON DIFFERENT COMBINATIONS OF SEGMENTATION BACKBONE NETWORKS WITH THE SAME WSL FRAMEWORK (ACDC DATASET).

Network	Dice↑	Acc↑	Pre↑	Sen↑	Spe↑	HD↓	ASD↓
3×U-Net	0.9141	0.9959	0.8958	0.9383	0.9927	8.0566	2.8806
3×Swin-UNet	0.7446	0.9791	0.6555	0.9142	0.9815	121.4224	51.4317
3×Mamba-UNet	0.9128	0.9958	0.8931	0.9395	0.9932	8.3386	2.7928
2×Swin-UNet+U-Net	0.8349	0.9678	0.7527	0.9371	0.9707	12.4786	4.3784
2×Mamba-UNet+U-Net	0.8646	0.9881	0.8018	0.9381	0.9902	10.6453	3.4873
2×Swin-UNet+Mamba-UNet	0.7646	0.9887	0.6452	0.9382	0.9897	70.6452	30.4387
2×Mamba-UNet+Swin-UNet	0.8648	0.9724	0.8026	0.9375	0.9760	20.7894	10.3467
2×U-Net+Swin-UNet	0.8750	0.9769	0.8206	0.9371	0.9807	5.8374	3.7564
2×U-Net+Mamba-UNet	0.9054	0.9704	0.8759	0.9369	0.9764	4.7834	1.7480
U-Net+Swin-UNet+Mamba-UNet	0.9171	0.9963	0.9095	0.9309	0.9920	3.9597	0.8810

TABLE IV
ABLATION STUDIES ON DIFFERENT COMBINATIONS OF SEGMENTATION BACKBONE NETWORKS WITH THE SAME WSL FRAMEWORK (GLAS DATASET).

Network	Dice↑	Acc↑	Pre↑	Sen↑	Spe↑	HD_{95} ↓	ASD↓
3×U-Net	0.9104	0.9956	0.8925	0.9342	0.9923	8.4127	2.9648
3×Swin-UNet	0.7438	0.9786	0.6521	0.9156	0.9812	122.3184	52.1049
3×Mamba-UNet	0.9133	0.9959	0.8941	0.9396	0.9931	8.2379	2.7863
2×Swin-UNet+U-Net	0.8337	0.9673	0.7516	0.9360	0.9701	12.6574	4.4279
2×Mamba-UNet+U-Net	0.8665	0.9884	0.8034	0.9388	0.9903	10.4876	3.5387
2×Swin-UNet+Mamba-UNet	0.7642	0.9879	0.6459	0.9378	0.9891	70.7815	30.2678
2×Mamba-UNet+Swin-UNet	0.8651	0.9729	0.8038	0.9374	0.9763	20.6935	10.3281
2×U-Net+Swin-UNet	0.8762	0.9772	0.8217	0.9376	0.9809	5.9038	3.7684
2×U-Net+Mamba-UNet	0.9047	0.9709	0.8764	0.9368	0.9762	4.7589	1.7362
U-Net+Swin-UNet+Mamba-UNet (ours)	0.9188	0.9964	0.9109	0.9318	0.9926	3.8945	0.8721

Mamba-based architecture provides a favorable balance between performance and computational efficiency, making it well suited for weakly supervised segmentation tasks where training resources are limited.

F. Qualitative Results

Figure 5 showcases the efficacy of our proposed method through three illustrative sample slices alongside their actual labels. These examples demonstrate how conventional pCE

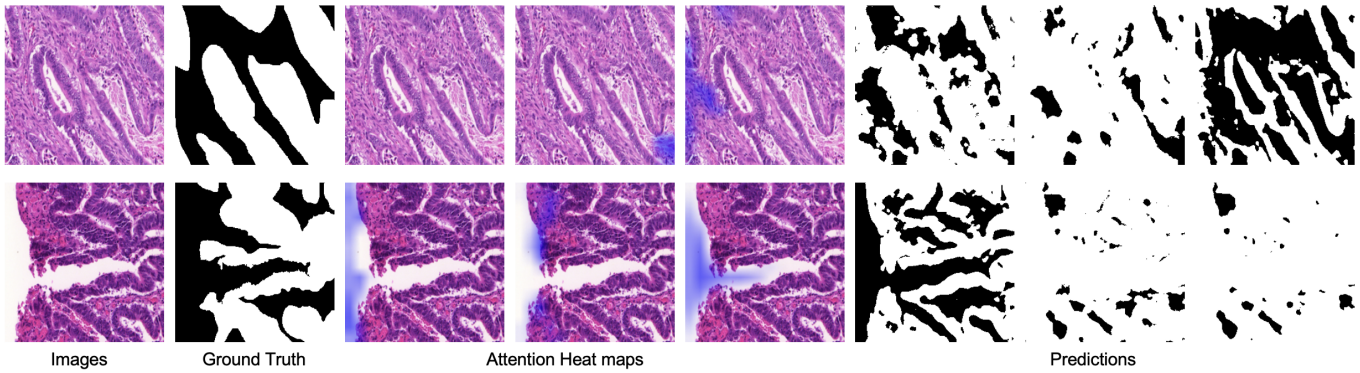


Fig. 7. The Illustration of two representative failure cases, showing the input images, ground truth segmentation, three corresponding attention maps, and three final predictions produced by three models.

TABLE V
MIXING STRATEGY FOR PSEUDO LABELS (DICE ON ACDC).

Scheme	Dice
Fixed ratio $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$	0.8048
Fixed ratio $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$	0.7853
Fixed ratio $(\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$	0.7954
Fixed ratio $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	0.8736
Dirichlet sampling $(\alpha, \beta, \gamma) \sim \mathrm{Dir}(1, 1, 1)$	<u>0.9171</u>

TABLE VI
COMPUTATIONAL EFFICIENCY COMPARISON OF U-NET, SWIN-UNET, AND MAMBA-UNET ON AN NVIDIA RTX 3090 GPU.

	U-Net	Swin-UNet	Mamba-UNet
FLOPs ($\times 10^9$)	1.97	8.69	**3.52**
Parameters ($\times 10^6$)	1.08	41.34	**15.48**
Memory (GB)	0.5	1.4	**1.1**
Training time (s/epoch)	42.8	88.3	**57.9**
Inference time (ms/image)	12.4	29.6	**17.3**

and USTM frameworks may lead to erroneous predictions, whereas our novel multi-model combination approach effectively addresses these issues, achieving superior segmentation outcomes.

As shown in Fig. 6, the proposed Weak-Mamba-UNet effectively highlights the most relevant anatomical regions through its attention mechanisms. The Grad-CAM visualizations demonstrate that the network focuses accurately on lesion boundaries and key structural areas, confirming its ability to capture both local details and global contextual information under weak supervision.

As depicted in Fig. 7, the identified failure cases typically occur in regions with ambiguous boundaries, low contrast, or overlapping anatomical structures. The attention maps indicate that the network occasionally misfocuses on irrelevant regions or fails to fully capture fine-grained contextual cues. These observations suggest potential directions for further improvement, such as incorporating uncertainty modeling or multi-scale context refinement.

G. Quantitative Results

The results of our quantitative comparison on the ACDC and GlaS dataset are detailed in Table I and Table II respectively, highlighting several key observations with the best-performing results <u>underscored</u>. Notably, WSL methods employing the Swin-UNet architecture (pCE-Swin-UNet and USTM-Swin-UNet) generally surpass those based on the U-Net framework (pCE-U-Net and USTM-U-Net). For instance, pCE-Swin-UNet exceeds pCE-U-Net in Dice and HD with scores of 0.7620 and 54.6531, respectively, underscoring the significance of employing advanced algorithms within the WSL framework. However, an optimized integration of multiple independent algorithms, as exhibited by Weak-Mamba-UNet, can yield even more impressive results.

To verify the reliability of the improvements, we performed a two-tailed paired t-test between our proposed Weak-Mamba-UNet and each baseline method on both Dice and HD_{95} metrics. All comparisons yielded p-values smaller than 1e-6, indicating that the observed performance gains are statistically significant. We further report the 95% confidence intervals (CIs) computed over five independent runs, which show narrow ranges across both datasets (see Tables I and II). These results confirm that the proposed framework consistently outperforms existing weakly supervised methods with high statistical confidence.

H. Ablation Study

The ablation studies presented in Table IV illustrate the contributions of the proposed WSL framework with different combinations of segmentation backbone networks. As can be seen from Table IV, the WSL framework consisting solely of Swin-UNet performs less well, which indicates that although the performance of the independent Swin-UNet algorithm is able to outperform that of U-Net, there is a lack of sufficient feature differentiation between the Multi-Swin-UNet models when combined within the same WSL framework. This result highlights that simply stacking similar architectures may not always yield better performance, as the networks might converge on similar feature representations, thereby limiting diversity and overall improvements in segmentation accuracy.

In contrast, it is worth noting that the Mamba-UNet model can significantly enhance the feature diversity among multiple Mamba-UNet instances by learning feature dependencies over longer distances. This allows Mamba-UNet to capture more global contextual information, which contributes to its strong performance across various metrics. This suggests that the

Visual Mamba architecture offers a unique advantage in capturing long-range dependencies that conventional CNN-based models may miss.

Table V presents an ablation on the weighting strategy used to combine predictions from the three heterogeneous backbones. When fixed ratios are applied, the performance varies depending on which model dominates, and the overall Dice remains below 0.88. By contrast, introducing stochastic perturbation through Dirichlet sampling—where $(\alpha, \beta, \gamma) \sim \mathrm{Dir}(1, 1, 1)$ are randomly sampled weights satisfying $\alpha + \beta + \gamma = 1$—yields a clear improvement (Dice $= 0.9171$). This randomization acts as a *model-side perturbation*, encouraging the network to maintain consistency under different mixtures of teacher predictions. Such perturbation-driven consistency regularization improves generalization and prevents the training from overfitting to any single backbone's bias.

Finally, our proposed WSL framework, which integrates different types of architectures—CNN, ViT, and Mamba—achieves optimal results on most segmentation metrics. This demonstrates that using multiple independent algorithms of distinct types, each with its own strengths, allows the networks to complement each other with different levels of feature information. By leveraging both local and global feature representations, the networks in the Weak-Mamba-UNet framework enhance overall segmentation performance, particularly in challenging weakly-supervised learning tasks where the available annotation signals are sparse or imprecise.

V. Conclusion

Weak-Mamba-UNet, by integrating the feature learning capabilities of CNN, ViT, and VMamba within a scribble-supervised learning framework, significantly reduces the costs and resources required for annotations. The multi-view cross-supervised learning approach enhances the adaptability of different network architectures, enabling them to mutually benefit from each other. Crucially, this study demonstrates the effectiveness of the novel Visual Mamba network architecture in medical image segmentation under limited signal supervision. The promising outcomes of this research not only highlight the network's high accuracy in segmentation tasks but also underscore the potential for broader applications in medical image analysis, particularly in settings where resources are limited.

Looking ahead, future work could explore the application of Weak-Mamba-UNet to a broader range of medical imaging modalities, such as CT and ultrasound, where sparse annotations are prevalent. Additionally, adapting this framework to handle multi-modal data or volumetric 3D imaging could further enhance its utility, particularly in complex medical segmentation tasks. Another promising direction is to extend the application of Weak-Mamba-UNet beyond medical image analysis to other domains, such as remote sensing or industrial monitoring, where weak annotations are also common. Exploring how the model performs on these more diverse datasets could unlock new opportunities for segmentation tasks across various fields, providing a generalizable and efficient solution where annotation resources are limited.

By refining and extending the use of Weak-Mamba-UNet, this research lays the groundwork for future innovations in both medical image segmentation and beyond, offering a scalable approach to address the growing demands for efficient annotation and segmentation in data-driven applications.

VI. Acknowledgment

We thank Chao Ma for providing valuable suggestions on writing and experimental design.

[1] Olivier 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):2514–2525, 2018.

[2] Léon Bottou. Stochastic gradient learning in neural networks. In *Proceedings of Neuro-Nîmes 91*, Nimes, France, 1991. EC2.

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[4] Hongjun Chen, Jinbao Wang, Hong Cai Chen, Xiantong Zhen, Feng Zheng, Rongrong Ji, and Ling Shao. Seminar learning for click-level weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6920–6929, 2021.

[5] Xiaokang Chen et al. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021.

[6] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021.

[9] W Dong-DongChen and ZH WeiGao. Tri-net for semi-supervised deep learning. In *Proceedings of twenty-seventh international joint conference on artificial intelligence*, pages 2014–2020, 2018.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Zishun Feng, Dong Nie, Li Wang, and Dinggang Shen. Semi-supervised learning for pelvic mr image segmentation based on multi-task residual fully convolutional networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 885–888. IEEE, 2018.

[12] Albert Gu. *Modeling Sequences with Structured State Spaces*. PhD thesis, Stanford University, 2023.

[13] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[14] Yuexing Han, Ruiqi Li, Bing Wang, Liheng Ruan, and Qiaochuan Chen. A pseudo-labeling based weakly supervised segmentation method for few-shot texture images. *Expert Systems with Applications*, 238:122110, 2024.

[15] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[16] Nabil Ibtehaz and M Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87, 2020.

[17] Rushi Jiao, Yichi Zhang, Le Ding, Bingsen Xue, Jicong Zhang, Rong Cai, and Cheng Jin. Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 2023.

[18] Boah Kim and Jong Chul Ye. Mumford–shah loss functional for image segmentation with deep learning. *IEEE Transactions on Image Processing*, 29:1856–1866, 2019.

[19] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652, 2021.

[20] Victor Lempitsky, Pushmeet Kohli, Carsten Rother, and Toby Sharp. Image segmentation with a bounding box prior. In *2009 IEEE 12th international conference on computer vision*, pages 277–284. IEEE, 2009.

[21] Zihan Li, Yuan Zheng, Xiangde Luo, Dandan Shan, and Qingqi Hong. Scribblevc: Scribble-supervised medical image segmentation with vision-class embedding. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3384–3393, 2023.

[22] Zihan Li, Yuan Zheng, Dandan Shan, Shuzhou Yang, Qingde Li, Beizhan Wang, Yuanting Zhang, Qingqi Hong, and Dinggang Shen. Scribformer: Transformer makes cnn work better for scribble-based medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024.

[23] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Lequan Yu, Yong Liang, Yizhou Yu, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-umamba†: Adapting mamba-based vision foundation models for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024.

[24] Xiaoming Liu, Quan Yuan, Yaozong Gao, Kelei He, Shuo Wang, Xiao Tang, Jinshan Tang, and Dinggang Shen. Weakly supervised segmentation of covid19 infection with scribble annotation on ct images. *Pattern recognition*, 122:108341, 2022.

[25] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[26] Ze Liu, Yutong Lin, et al. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[28] Wenfeng Luo, Meng Yang, and Weishi Zheng. Weakly-supervised semantic segmentation with saliency and incremental supervision updating. *Pattern Recognition*, 115:107858, 2021.

[29] Xiangde Luo et al. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. *arXiv preprint arXiv:2112.04894*, 2021.

[30] Xiangde Luo et al. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *MIDL*, 2022.

[31] Xiangde Luo, Minhao Hu, Wenjun Liao, Shuwei Zhai, Tao Song, Guotai Wang, and Shaoting 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*, pages 528–538. Springer, 2022.

[32] Xiangde Luo, Zihan Li, Shaoting Zhang, Wenjun Liao, and Guotai Wang. Rethinking abdominal organ segmentation (raos) in the clinical scenario: A robustness evaluation benchmark with challenging cases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 531–541. Springer, 2024.

[33] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[34] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013.

[35] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[36] Anton Obukhov et al. Gated crf loss for weakly supervised semantic image segmentation. *arXiv preprint arXiv:1906.04651*, 2019.

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

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

[39] Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

[40] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.

[41] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1818–1827, 2018.

[42] Antti Tarvainen and Harri 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*, pages 1195–1204, 2017.

[43] Gabriele Valvano, Andrea Leo, and Sotirios A Tsaftaris. Learning to segment from scribbles using multi-scale adversarial attention gates. *IEEE Transactions on Medical Imaging*, 40(8):1990–2001, 2021.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[45] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *International Joint Conference on Artificial Intelligence*, pages 3635–3641, 2019.

[46] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[47] Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong. Mixed transformer u-net for medical image segmentation. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2390–2394. IEEE, 2022.

[48] Jinhong Wang, Jintai Chen, Danny Chen, and Jian Wu. Lkm-unet: Large kernel vision mamba unet for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 360–370. Springer, 2024.

[49] Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397, 2023.

[50] Ziyang Wang, Tianxiang Chen, Zi Ye, Yiyuan Ge, Zhihao Chen, Jiabao Li, and Yifan Zhao. S4roboformer: Scribble-supervised surgical robotic segmentation transformer via augmented consistency training. *IEEE Transactions on Medical Robotics and Bionics*, 2025.

[51] Ziyang Wang 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] Ziyang Wang et al. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*, 2024.

[53] Ziyang Wang and Congying 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*, pages 870–879, 2023.

[54] Ziyang Wang and Chao Ma. Semi-mamba-unet: Pixel-level contrastive cross-supervised visual mamba-based unet for semi-supervised medical image segmentation. *arXiv preprint arXiv:2402.07245*, 2024.

[55] Ziyang Wang, Meiwen Su, Jian-Qing Zheng, and Yang Liu. Densely connected swin-unet for multiscale information aggregation in medical image segmentation. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 940–944. IEEE, 2023.

[56] Ziyang Wang and Irina 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)*, 2021.

[57] Ziyang Wang and Irina Voiculescu. Triple-view feature learning for medical image segmentation. In *MICCAI Workshop on Resource-Efficient Medical Image Analysis*, pages 42–54. Springer, 2022.

[58] Ziyang Wang and Irina 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*, pages 181–190. Springer, 2023.

[59] Ziyang Wang and Irina Voiculescu. Weakly supervised medical image segmentation through dense combinations of dense pseudo-labels. In *MICCAI Workshop on Data Engineering in Medical Imaging*, pages 1–10. Springer, 2023.

[60] Ziyang Wang, Haodong Zhang, and Yang Liu. Weakly-supervised self-ensembling vision transformer for mri cardiac segmentation. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, pages 101–102. IEEE, 2023.

[61] Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. H-vmunet: High-order vision mamba unet for medical image segmentation. *Neurocomputing*, page 129447, 2025.

[62] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Seg-

mamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560*, 2024.

[63] Rui Xu, Shu Yang, Yihui Wang, Bo Du, and Hao Chen. A survey on vision mamba: Models, applications and challenges. *arXiv preprint arXiv:2404.18861*, 2024.

[64] Zhongxing Xu, Feilong Tang, Zhe Chen, Zheng Zhou, Weishan Wu, Yuyao Yang, Yu Liang, Jiyu Jiang, Xuyue Cai, and Jionglong Su. Polyp-mamba: Polyp segmentation with visual mamba. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 510–521. Springer, 2024.

[65] Zihong Xu and Ziyang Wang. Mcv-unet: a modified convolution & transformer hybrid encoder-decoder network with multi-scale information fusion for ultrasound image semantic segmentation. *PeerJ Computer Science*, 10:e2146, 2024.

[66] Xiangyi Yan et al. After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3971–3981, 2022.

[67] Yijun Yang, Zhaohu Xing, and Lei Zhu. Vivim: a video vision mamba for medical video object segmentation. *arXiv preprint arXiv:2401.14168*, 2024.

[68] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[69] Zi Ye and Tianxiang Chen. P-mamba: Marrying perona malik diffusion with mamba for efficient pediatric echocardiographic left ventricular segmentation. *arXiv preprint arXiv:2402.08506*, 2024.

[70] Hanwei Zhang, Ying Zhu, Dan Wang, Lijun Zhang, Tianxiang Chen, Ziyang Wang, and Zi Ye. A survey on visual mamba. *Applied Sciences*, 14(13):5683, 2024.

[71] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International conference on medical image computing and computer-assisted intervention*, pages 408–416. Springer, 2017.

[72] Yichi Zhang, Lin Yuan, Yujia Wang, and Jicong Zhang. Sau-net: efficient 3d spine mri segmentation using inter-slice attention. In *Medical Imaging With Deep Learning*, pages 903–913. PMLR, 2020.

[73] Zhengdong Zhang, Shuai Li, Ziyang Wang, and Yun Lu. A novel and efficient tumor detection framework for pancreatic cancer via ct images. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1160–1164. IEEE, 2020.

[74] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 2023.

[75] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming 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*, pages 3–11. Springer, 2018.

[76] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.