Biomedical Signal Processing and Control

PCMamba: Parallel Convolution-Mamba Network for Medical Image Segmentation

--Manuscript Draft--

Manuscript Number:	BSPC-D-25-00813
Article Type:	Research Paper
Keywords:	State space model (SSM); Reparameterization; Large Kernel Convolution; Wavelet Transform
Corresponding Author:	Fei Yang Shandong University CHINA
First Author:	JiaCheng Liu
Order of Authors:	JiaCheng Liu
	Liquan Dong
	Bo Chen
	Jiahui Shi
	Weigang Lu
	Shugang Zhang
	Fei Yang
	Dong Li
	Lei Zhang
Abstract:	To address the challenge of balancing coarse-grained and fine-grained segmentation in medical image segmentation, we propose a comprehensive solution that integrates State space mode(SSMs) with innovative convolutional methods for superior local and global segmentation performance. In the SSMs, we incorporate DSConv and wavelet transform to capture local spatial features and frequency-domain characteristics. Additionally, we design two convolution strategies——DiscWideFusion and RepLKWideFusion——to enhance the capability for fine-grained segmentation. DiscWideFusion expands the receptive field through dilated convolutions, effectively capturing multi-scale local features while reducing parameter count and computational complexity. Meanwhile, RepLKWideFusion employs a combination of small-kernel and large-kernel convolutions, leveraging FFT-based acceleration for large-kernel convolutions to improve sensitivity to intricate structures. To balance local and global segmentation effectiveness, we propose the Metric-Adaptive Loss (MAL) function, which adaptively integrates multi-level key metrics to dynamically gate segmentation task characteristics. This approach significantly improves training efficiency and overall model performance. Experimental results demonstrate that our model achieves state-of-the-art (SOTA) performance across four modalities and nine datasets. Notably, on the ACDC dataset, the model achieves an average Dice coefficient of 93.77%, setting a new benchmark for segmentation accuracy.

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Graphical Abstract

PCMamba: Parallel Convolution-Mamba Network for Medical Image Segmentation

Jiacheng Liu, Liquan Dong, Bo Chen, Jiahui Shi, Weigang Lu, Shugang Zhang, Fei Yang, Dong Li, Lei Zhang

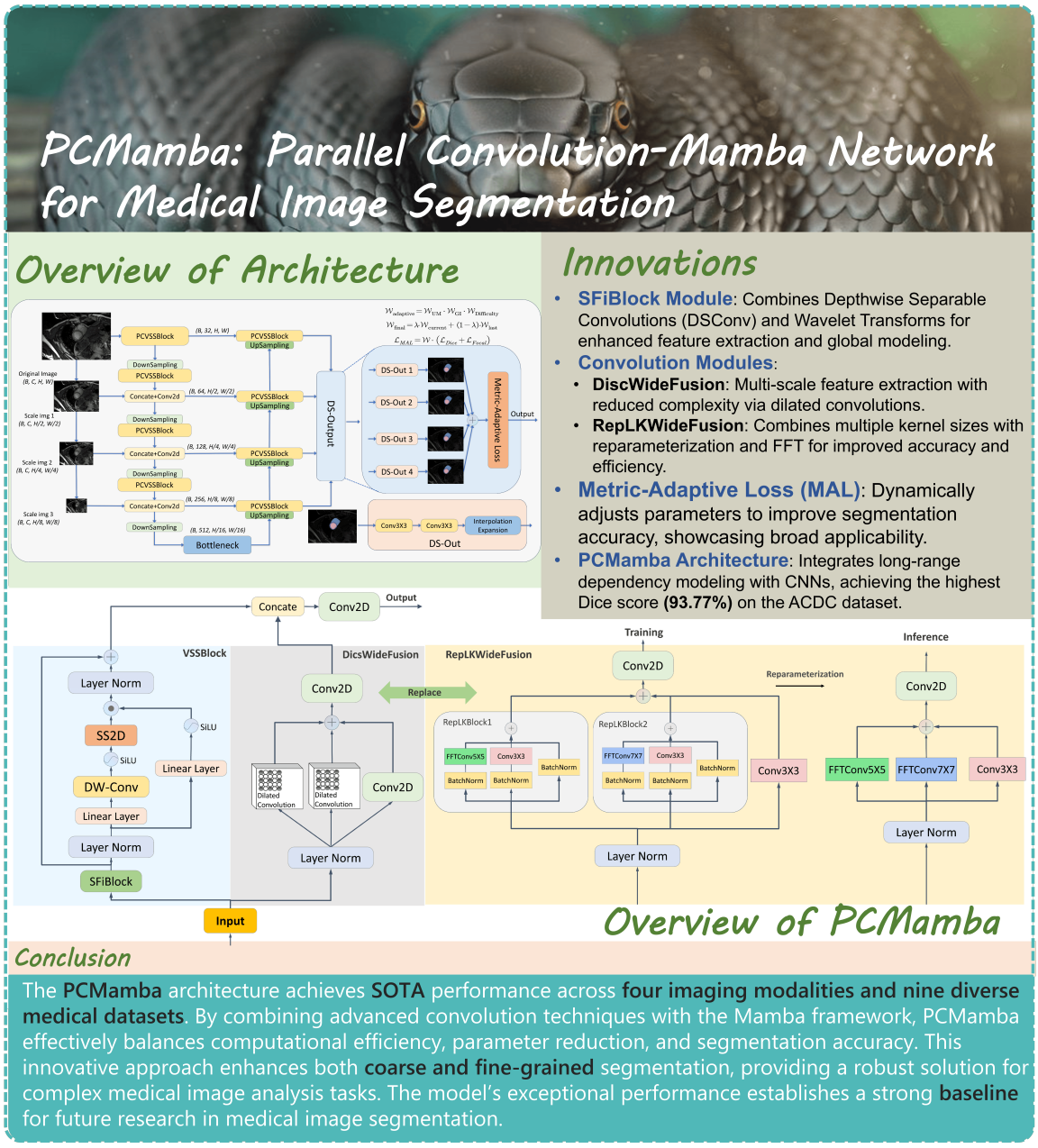

Figure 1: Graphical Abstract: A concise visual representation of the PCMamba architecture.

PCMamba: Parallel Convolution-Mamba Network for Medical Image Segmentation

Jiacheng Liu[a], Liquan Dong[c], Bo Chen[a], Jiahui Shi[a], Weigang Lu[d,e], Shugang Zhang[e], Fei Yang[a,b,*], Dong Li[a,b], Lei Zhang[f]

[a]School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, 264209, China
[b]Shandong Key Laboratory of Intelligent Electronic Packaging Testing and Application, Shandong University, Weihai, 264209, China
[c]Department of General Practice, Weihai Municipal Hospital, Cheeloo College of Medicine, Shandong University, Weihai, 264200, China
[d]Department of Educational Technology, Ocean University of China, Qingdao, Qingdao, 266100, China
[e]Department of Computer Science and Technology, Ocean University of China, Qingdao, Qingdao, 266100, China
[f]Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland, Baltimore, MD 21201, Maryland, USA

Abstract

To address the challenge of balancing coarse-grained and fine-grained segmentation in medical image segmentation, we propose a comprehensive solution that integrates State space mode(SSMs) with innovative convolutional methods for superior local and global segmentation performance. In the SSMs, we incorporate DSConv and wavelet transform to capture local spatial features and frequency-domain characteristics. Additionally, we design two convolution strategies——**DiscWideFusion** and **RepLKWideFusion**——to enhance the capability for fine-grained segmentation. DiscWideFusion expands the receptive field through dilated convolutions, effectively capturing multi-scale local features while reducing parameter count and computational complexity. Meanwhile, RepLKWideFusion employs a combination of small-kernel and large-kernel convolutions, leveraging FFT-based acceleration for large-kernel convolutions to improve sensitivity to intricate structures. To balance local and global segmentation effectiveness, we propose the **Metric-Adaptive Loss (MAL)** function, which adaptively integrates multi-level key metrics to dynamically gate segmentation task characteristics. This approach significantly improves training efficiency and overall model performance. Experimental results demonstrate that our model achieves state-of-the-art (SOTA) performance across four modalities and nine datasets. Notably, on the ACDC dataset, the model achieves an average Dice coefficient of **93.77%**, setting a new benchmark for segmentation accuracy.

Keywords:
State space model (SSM), Reparameterization, Large Kernel Convolution, Wavelet Transform

1. Introduction

Medical semantic segmentation is a critical subfield of image segmentation that has gained considerable attention due to its impact on clinical diagnostics and treatment planning. Over the past decade, advancements in deep learning have driven substantial progress in segmentation tasks, particularly in organ segmentation, tumor detection, and vascular segmentation, where precise delineation of anatomical structures is vital. However, medical image segmentation remains a highly challenging task due to inherent difficulties in medical imaging data, including the complexity of anatomical structures, variations in pathological conditions, and low signal-to-noise ratios.

One of the primary challenges in medical segmentation is balancing coarse- and fine-grained segmentation. **Coarse-grained** segmentation typically involves the identification of large anatomical structures, such as organs, with the goal of rapid and accurate identification and segmentation. However, in medical images, complex backgrounds, inter-organ variations, and noise often complicate this task, leading to models struggling to accurately delineate these large structures, particularly in the presence of pathological abnormalities or low-quality imaging.

On the other hand, **fine-grained** segmentation focuses on the detection and segmentation of smaller, more complex structures, such as tumors, lesions, or fine anatomical boundaries. This task demands that the model capture subtle details with a high degree of precision. However, fine-grained segmentation in medical images presents unique challenges because of the high resolution of the images and the complex relationships between various fine structures. Moreover, the fine details of these structures can often be disrupted by noise, small artifacts, or ambiguous boundaries, potentially leading to overfitting or poor generalization, especially when applying models to new or unseen datasets. The challenge is further compounded by the necessity for the model to generalize effectively across varying imaging conditions, patient populations, and pathological scenarios.

A key challenge in current methods is the difficulty of simultaneously addressing both coarse-grained and fine-grained segmentation tasks. While some models excel in specific tasks,

*Corresponding author
Email addresses: 202200800027@mail.sdu.edu.cn (Jiacheng Liu), dongliquan1@163.com (Liquan Dong), 202137544@mail.sdu.edu.cn (Bo Chen), 202417561@mail.sdu.edu.cn (Jiahui Shi), luweigang@ouc.edu.cn (Weigang Lu), zsg@ouc.edu.cn (Shugang Zhang), feiyang@sdu.edu.cn (Fei Yang), dongli@sdu.edu.cn (Dong Li), LeiZhang@som.umaryland.edu (Lei Zhang)

they often underperform in others. For instance, models focused on large structures may fail to capture fine boundaries, while models optimized for details may overlook global context or larger anatomical features. This limitation affects the effectiveness of many methods in clinical applications requiring both segmentation types.

Therefore, the core challenge is to develop a framework capable of efficiently handling both coarse and fine-grained segmentation tasks. The model must maintain strong contextual awareness for large structures, be sensitive to fine details, and dynamically balance these tasks, adapting to the varying complexities of different regions within the image.

An example illustrating this dual challenge is the ACDC cardiac segmentation dataset[1]. Models must accurately segment large regions such as the left ventricle, right ventricle, and myocardium while delineating fine boundaries within these structures, even under pathological conditions like myocardial infarction or heart failure. The structural variations in these regions underscore the need for models capable of balancing global and fine-grained segmentation, which is critical for addressing complex medical imaging challenges.

Developing models capable of handling both coarse and fine-grained segmentation tasks is essential for advancing medical imaging. Successfully addressing this challenge will not only enhance segmentation performance but also broaden clinical applications, ultimately improving patient diagnosis and treatment outcomes. In this study, we introduce the **PCMamba** architecture, a novel framework designed to dynamically balance multi-scale attention for precise anatomical segmentation. Our key contributions are as follows:

(1) We designed the **SFiBlock** module to combine DSConv with Wavelet Transforms to effectively extract both local spatial and frequency domain features, thereby substantially improving the feature extraction and global modeling capabilities of SSMs.

(2)We introduce two novel convolution modules: **DiscWideFusion**, designed to capture multi-scale features via dilated convolutions, reducing both model parameters and computational load; and **RepLKWideFusion**, which optimizes accuracy and efficiency by combining kernels of various sizes. This is further accelerated by leveraging reparameterization to merge kernels during inference, and by incorporating FFT for efficient large-kernel convolution computation, achieving superior performance in both speed and accuracy.

(3) We present the **PCMamba** architecture, a novel paradigm that combines Mamba's long-range dependency modeling capabilities with CNNs, demonstrating superior performance in both coarse and fine-grained segmentation tasks. Notably, on the ACDC dataset, we achieved the highest recorded Dice score to date (93.77%).

(4) We propose the **Metric-Adaptive Loss (MAL)** function, which adaptively adjusts key parameters in response to the model's performance across various tasks, significantly enhancing the segmentation accuracy of the PCMamba model and other segmentation frameworks, demonstrating its broad applicability and flexibility.

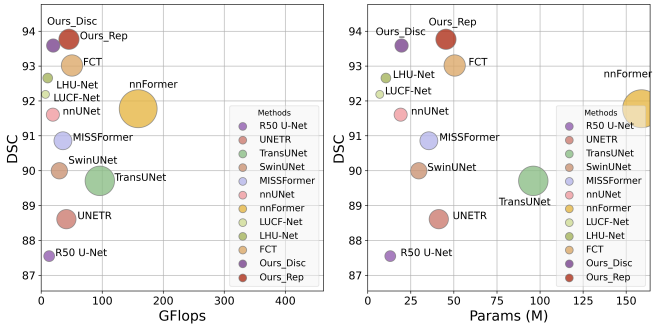

Figure 2: Performance Comparison of Our Models (Ours_Disc, Ours_Rep) with State-of-the-Art Methods on the ACDC Dataset: Evaluation in Terms of GFlops and Parameters

2. Related work

2.1. Methods Based on CNN

Convolutional Neural Networks (CNNs) have gained prominence in computer vision tasks due to their ability to capture local spatial information. The ResNet architecture[2] mitigates the gradient vanishing problem through the use of skip connections, facilitating the training of deeper networks. ConvNeXt[3] and its subsequent version, ConvNeXt V2[4], further refine network structures to enhance performance.

In medical image segmentation, CNNs are frequently integrated with decoders, with U-Net being the predominant model. Variants such as RA-UNet[5] enhance accuracy for small organ segmentation by incorporating residual and attention mechanisms. The nnU-Net[6] further improves segmentation precision through automated hyperparameter optimization.

2.2. Vision Transformers and Their Variants

Inspired by the success of attention mechanisms in natural language processing, researchers have incorporated self-attention into computer vision, leading to the development of models such as Vision Transformer (ViT[7]), Swin Transformer[8], and PVT[9]. ViT strikes an effective balance between speed and accuracy by partitioning images into fixed-size patches, achieving performance comparable to top CNN models on ImageNet[10]. However, a key limitation of ViT is its dependence on large datasets, which increases the risk of overfitting when applied to smaller datasets. Additionally, the self-attention mechanism requires significant computational resources and memory, especially for high-resolution images, which has spurred various optimization strategies.

2.3. Combination of Transformer and CNNs

U-Net-based architectures, such as TransUNet[11], Swin-Unet[12], Medical Transformer[13], and MPU-Net[14], leverage the combined strengths of Transformers and CNNs. These models typically employ Transformers, including ViT and Swin Transformer, as core encoders to incorporate visual attention during downsampling for feature extraction, followed by decoders for upsampling to produce segmentation results. However, the uniformity in the size of feature extraction units across

both the encoder and decoder underscores the necessity for optimizing Transformer block utilization and refining network structures for effective upsampling and downsampling.

Swin-Unet overcomes the limitations of CNNs in effectively capturing global information during feature extraction by leveraging Transformer-based architectures. It employs Swin-Transformer Blocks as the primary feature extraction units within both the encoder and decoder, using linear layers to learn implicit spatial feature representations. This approach strengthens the model's capability to capture global context, thereby improving segmentation accuracy and achieving outstanding performance across a range of medical image segmentation tasks. PVT is designed to reduce memory consumption while maintaining high performance. It utilizes specific combinations of CNN modules for downsampling, resulting in pyramid-like feature maps that not only minimize memory usage but also enhance the diversity of extracted features.

2.4. Mamba for Medical Image Segmentation

State Space Models (SSMs)[15] are mathematical frameworks for describing the behavior of dynamic systems. However, early models were constrained by high computational and memory demands, limiting their widespread application. The Structured State Space Sequence Model (S4) improved the efficiency of convolution kernel computations by leveraging a Normal Plus Low-Rank (NPLR) representation and the Woodbury identity. These advancements enabled the integration of SSMs into end-to-end neural network architectures, driving the development of multiple SSM variants.

Despite these advancements, fixed sequence transformations in SSMs still limit their ability to perform context-based reasoning. To overcome this limitation, Mamba (Selective SSM)[16] introduced time-varying parameters and hardware-aware algorithms, significantly improving both training and inference efficiency, particularly in long-sequence modeling.

Following the success of SSMs, researchers began applying them to computer vision tasks, similar to the adaptation process of Transformers. VMamba[17] incorporated cross-scan modules with Mamba to address the directional sensitivity between 1D sequences and 2D images, while Vim introduced a bidirectional SSM for global context modeling using position embeddings.

In medical imaging, U-Mamba[18] integrated Mamba layers into the nnUNet encoder, improving the CNN's ability to model long-range dependencies. Meanwhile, VM-UNet[19] established a new benchmark as the first purely SSM-based medical image segmentation model. Additionally, Swin-UMamba[20] showcased the benefits of ImageNet pre-training for enhancing Mamba's performance in medical segmentation. LMa-UNet[21] introduced a large-window mechanism to improve context capture, while Weak-Mamba-UNet[22] proposed a weakly-supervised learning framework that integrates the feature learning capabilities of CNNs, ViTs, and VMamba, significantly reducing annotation costs and resource demands.

3. Proposed Method

3.1. Preliminaries

In contemporary approaches based on state space models (SSMs), such as the Structured State Space Sequence Model (S4) and Mamba, both rely on a classic continuous system. This system maps a one-dimensional input function or sequence $x(t) \in \mathbb{R}$ through an intermediate hidden state $h(t) \in \mathbb{R}^N$ to an output $y(t) \in \mathbb{R}$. This process can be described by a linear ordinary differential equation (ODE):

$$\begin{aligned} \mathbf{h}'(t) &= \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{h}(t) \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ represents the state matrix, while $\mathbf{B} \in \mathbb{R}^{N\times 1}$ and $\mathbf{C} \in \mathbb{R}^{N\times 1}$ denote the projection parameters.

To make these continuous systems more suitable for deep learning scenarios, S4 and Mamba adopt a discretization strategy. They introduce a time-scale parameter Δ and transform $\mathbf{A}$ and $\mathbf{B}$ into discrete parameters $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ using a fixed discretization rule. Typically, the zero-order hold (ZOH) is used as the discretization rule, defined as follows:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}) \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \end{aligned} \tag{2}$$

After discretization, SSM-based models can be computed in two ways: linear recurrence or global convolution, defined as equations 3 and 4, respectively.

$$\begin{aligned} h'(t) &= \overline{\mathbf{A}}h(t) + \overline{\mathbf{B}}x(t) \\ y(t) &= \mathbf{C}h(t) \end{aligned} \tag{3}$$

For the global convolution method, we have:

$$\begin{aligned} \overline{\mathbf{K}} &= (\overline{\mathbf{C}\mathbf{B}}, \overline{\mathbf{C}\mathbf{A}\mathbf{B}}, \ldots, \overline{\mathbf{C}\mathbf{A}^{L-1}\mathbf{B}}) \\ y &= x * \overline{\mathbf{K}} \end{aligned} \tag{4}$$

where $\overline{\mathbf{K}} \in \mathbb{R}^L$ represents a structured convolutional kernel, and L denotes the length of the input sequence x.

3.2. WideFusion Module

We designed two convolution modules: **DiscWideFusion** and **RepLKWideFusion**, each tailored for specific optimizations in convolutional operations.

DiscWideFusion leverages three discrete convolutions with varying dilation rates to effectively expand the receptive field, enabling precise capture of multi-scale local features. Unlike traditional large-kernel approaches, this design achieves remarkable segmentation performance while significantly reducing parameter count and computational overhead by utilizing variable dilation rates.

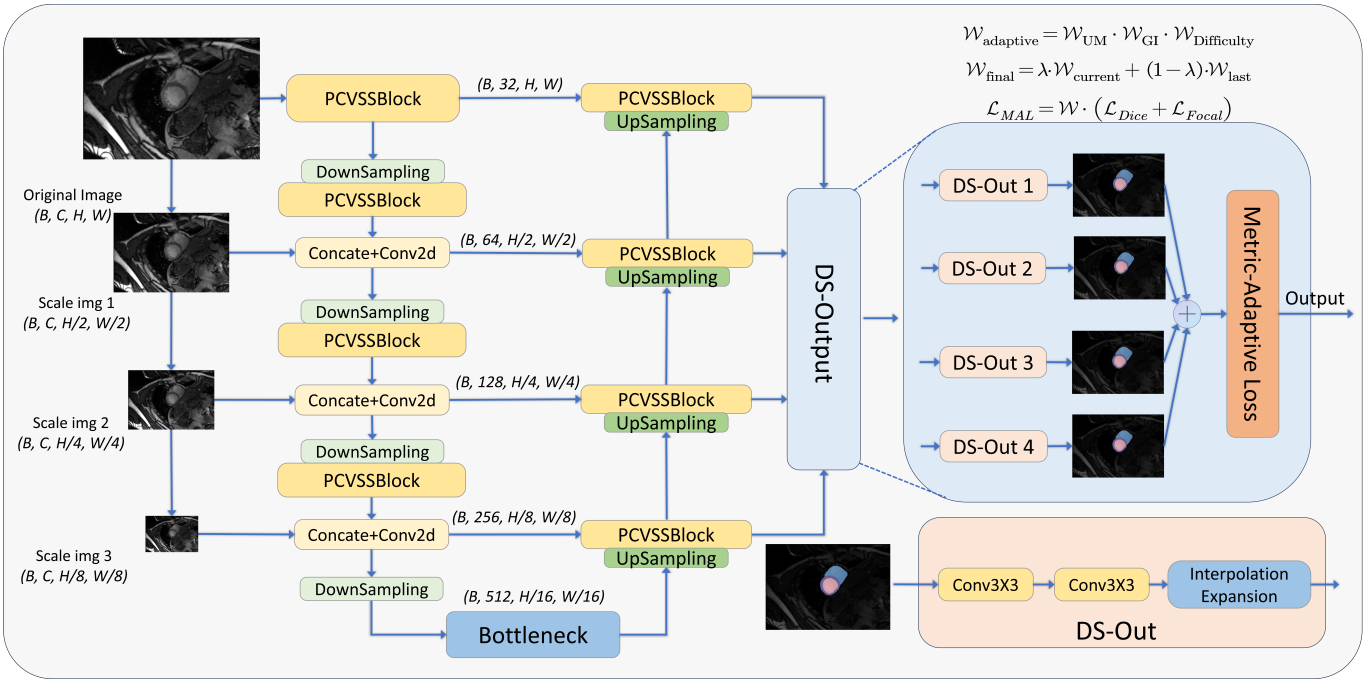

Figure 3: Architecture Overview: Multi-level inputs extract multi-scale features. The DS-Up module restores the original scale for loss computation, guided by a Metric-Adaptive loss function.

In the **DiscWideFusion** module, the convolution operation can be described as follows:

$$\mathbf{X_{out}} = \text{Conv}_3(\mathbf{X}) + \text{Disc}_5(\mathbf{X}) + \text{Disc}_7(\mathbf{X}) \tag{5}$$

For an input feature map of size $H \times W$, input channels C_{in}, output channels C_{out}, and a kernel size of $K \times K$, the computational complexity (FLOPs) of a standard convolution is is given by $\mathbf{F}_{\text{std}} = H \times W \times K^2 \times C_{\text{in}} \times C_{\text{out}}$, and the number of parameters is $\mathbf{P}_{\text{std}} = K^2 \times C_{\text{in}} \times C_{\text{out}}$.

The introduction of dilated convolutions significantly expands the receptive field without substantially increasing computational complexity (FLOPs). Let d be the dilation rate, and the effective coverage of the kernel is $K' = (K-1) \cdot d + 1$. Therefore, the computational complexity of dilated convolutions is $\mathbf{F}_{\text{dil}} = H \times W \times \frac{K'^2}{d^2} \times C_{\text{in}} \times C_{\text{out}}$. The use of d reduces the number of elements involved in the computation, maintaining a lower computational cost compared to direct use of larger kernels.

In terms of the number of parameters, dilated convolutions have the same parameter count as standard convolutions, $\mathbf{P}_{\text{dil}} = K^2 \times C_{\text{in}} \times C_{\text{out}}$. This indicates that dilated convolutions extend the receptive field and enhance feature extraction capabilities without incurring additional parameter costs, unlike direct use of larger kernels that require extra parameters.

However, at high dilation rates, the grid effect can lead to uneven coverage and reduced long-range information correlation, limiting segmentation accuracy. While dilated convolutions can effectively expand the receptive field without increasing parameters, the introduced sparsity can reduce feature extraction capabilities. To overcome these limitations, we drew inspiration from RepLKNet[23] (CVPR 2022) and explored the use of large-kernel convolutions. Large-kernel convolutions provide continuous coverage and better maintain information correlation. This shift not only avoids the drawbacks of uneven sampling but also significantly enhances feature extraction performance.

The inevitable issue with large kernel convolutions is their computational inefficiency. To address the computational burden of large convolution kernels, we propose the **RepLKWideFusion** module, which combines reparameterization and FFT for accelerated convolution, and optimizes computational overhead through multi-scale feature fusion and kernel merging. This method surpasses traditional large kernel convolutions in terms of detail retention and feature representation, significantly improving computational efficiency.

Traditional spatial-domain convolution is computationally expensive, particularly when the kernel size increases. Let the input image be I and the convolution kernel be K, , and the convolution output be O. The spatial-domain convolution can then be expressed as:

$$O(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i-m, j-n) K(m,n) \tag{6}$$

This 2D sliding window calculation in the spatial domain quickly becomes computationally expensive as the image and kernel sizes grow. To accelerate convolution, Fourier transform provides an efficient solution.

According to the convolution theorem, the spatial-domain convolution can be transformed into a pointwise multiplication

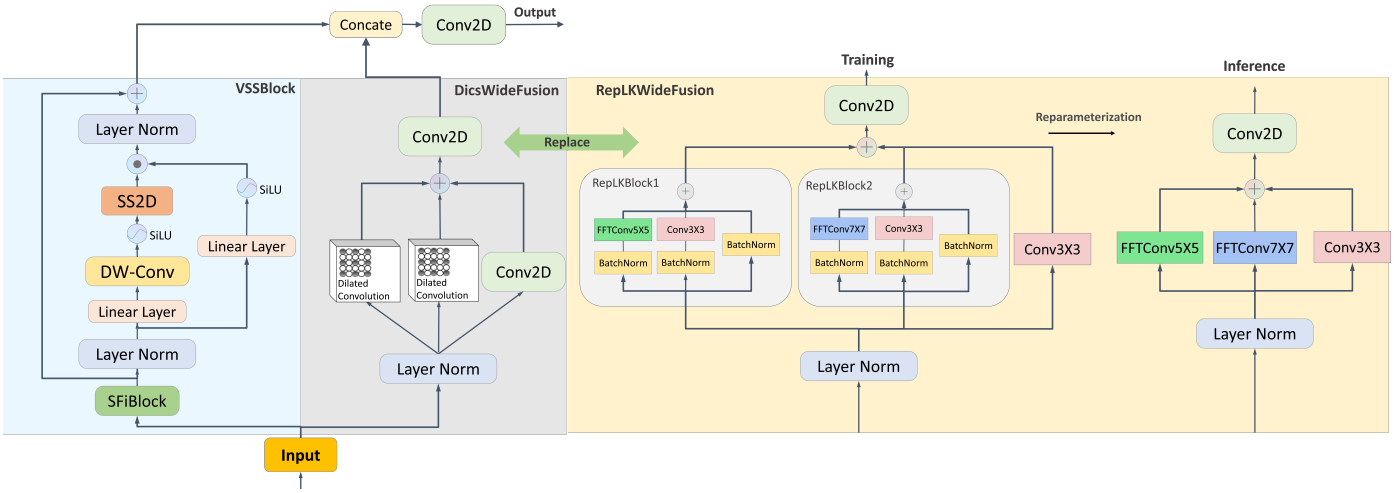

Figure 4: Overview of PCMamba, We propose a parallel architecture combining VSSBlock and WideFusion, highlighting the re-parameterization strategy of the WideFusion module. **DiscWideFusion** captures multi-scale features using two discrete convolutions and one standard convolution in parallel; **RepLKWideFusion** fuses convolution kernels with BatchNorm through reparameterization to reduce computational overhead.

operation in the frequency domain:

$$O = \mathcal{F}^{-1}\left(\mathcal{F}(I) \odot \mathcal{F}(K)\right) \tag{7}$$

Here, $\mathcal{F}(I)$ and $\mathcal{F}(K)$ denote the frequency domain representations of the input image and the convolution kernel, respectively. The element-wise multiplication, denoted by $\odot$, is performed in the frequency domain. The resulting product is then transformed back to the spatial domain using the inverse Fourier transform (IFFT), as expressed in the following equation:

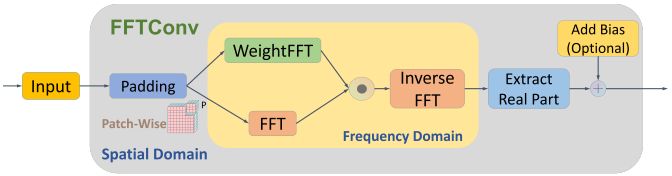

Figure 5: FFTConv: Accelerating Large Kernel Convolution by Transforming Spatial Convolution to Frequency Domain Pointwise Multiplication Using Fourier Transform.

The Fourier transform simplifies the convolution operation by converting it from addition in the spatial domain to element-wise multiplication in the frequency domain, thereby significantly reducing computational complexity. In the spatial domain, the traditional computational complexity of convolution is $O(C_{\text{in}}C_{\text{out}}K^2HW)$, where C_{in} and C_{out} represent the number of input and output channels, K is the kernel size, and H and W are the height and width of the input image. By leveraging the Fourier transform to accelerate convolution, the complexity is reduced to $O(C_{\text{in}}C_{\text{out}}(HW)\log(HW))$. This reduction is particularly advantageous for large kernel sizes, leading to substantial speedup in computation.

To further enhance efficiency, we employ a patch-wise computation strategy. Instead of applying the Fourier transform to the entire image globally, the input image is divided into smaller patches. The Fourier transform is then applied to each patch individually, and convolution is performed independently within the frequency domain. This patch-wise approach significantly reduces both computational and memory overhead compared to a global Fourier transform, making it especially effective for processing high-resolution images. By combining Fourier acceleration with patch-wise computation, this method achieves a marked improvement in efficiency while maintaining scalability.

To further optimize computational efficiency during inference, the **RepLKWideFusion** module introduces reparameterization of convolution kernels. The core idea is to merge multiple convolution kernels of different sizes, thereby reducing redundant computations and memory usage. Specifically, during training, the model uses convolution kernels of various sizes (e.g., 3×3, 5×5, 7×7) for multi-scale feature extraction. During inference, the reparameterization technique merges these kernels into a single equivalent kernel, thus reducing the computational burden while preserving the advantages of multi-scale feature fusion.

Suppose we have two convolution kernels, $\mathbf{W}_{\text{large}}$ and $\mathbf{W}_{\text{small}}$, with sizes k_{large} and k_{small}, respectively. The reparameterization process involves padding the small kernel to match the size of the large kernel and then combining them into a single equivalent convolution kernel W_{eq}, as follows:

$$\mathbf{W}_{\text{eq}} = \mathbf{W}_{\text{large}} + \text{Pad}(\mathbf{W}_{\text{small}}, \frac{k_{\text{large}} - k_{\text{small}}}{2}) \tag{8}$$

Where $\text{Pad}(\mathbf{W}_{\text{small}}, p)$ represents padding the small kernel to match the size of the large kernel. This strategy not only reduces computational complexity but also ensures effective multi-scale information fusion, which further improves the granularity and expressive power of feature extraction.

Additionally, Batch Normalization (BN) parameters are

tightly coupled with convolution kernels. To preserve the model's performance during inference, we fuse the scaling (γ) and shifting (β) parameters with the kernels. The fused convolution kernel $\mathbf{W}_{\text{eq}}$ and bias $\mathbf{b}_{\text{eq}}$ are computed as:

$$\mathbf{W}_{\text{eq}} = \frac{\mathbf{W}_{\text{conv}} \cdot \gamma}{\sqrt{\sigma^2 + \epsilon}}, \quad \mathbf{b}_{\text{eq}} = \beta - \frac{\gamma \cdot \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{9}$$

Where μ and σ^2 are the mean and variance of BN, and ϵ prevents division by zero. This fusion reduces computational cost and ensures stability and accuracy during inference.

During inference, the **RepLKWideFusion** module combines reparameterization and FFT acceleration to reduce computational and memory overhead. Multiple convolution kernels are merged into a single equivalent kernel during training, and the convolution is accelerated using FFT. This approach significantly reduces the computational burden while retaining multi-scale feature fusion for efficient handling of complex features.

Specifically, the fused kernel $\mathbf{W}_{\text{eq}}$ is used for convolution, with FFT-accelerated pointwise multiplication speeding up the process. Compared to traditional convolution, FFT acceleration reduces redundant computations and memory usage, enabling faster inference while preserving multi-scale feature representation.

3.3. VSSBlock

The SS2D module is based on the continuous state space model (S4) and focuses on capturing long-range dependencies in images. SSMs are designed to model sequential data and capture long-term dependencies. SS2D employs a precise discretization for long-range modeling, reducing computational complexity compared to traditional self-attention mechanisms while increasing efficiency. It also integrates the DropPath mechanism and residual connections to enhance robustness and feature learning.

In the Mamba framework, SS2D primarily focuses on modeling long-range dependencies, demonstrating excellent performance in sequential tasks. However, images exhibit strong spatial-local features, and relying solely on long-range modeling is insufficient for effectively capturing fine-grained details. This issue is particularly evident in ultrasound images, which are prone to significant noise, where long-range modeling alone tends to be less effective. To address this limitation, we introduce DSConv and Wavelet Transform. DSConv is primarily responsible for extracting local spatial features $\mathcal{F}_{spatial}$, while the Wavelet Transform captures frequency-domain features $\mathcal{F}_{wavelet}$ through multi-scale decomposition, further enhancing the ability to capture image details.

The convolution operation is formulated as follows:

$$\text{DSConv}(x) = \text{DWConv}(x) + \text{PWConv}(x) \tag{10}$$

Here, DWConv operates on each channel of the input, capturing local spatial information across the $\mathbb{R}^{C\times H\times W}$ dimensions, while PWConv performs a 1x1 convolution to mix the information across channels. After this step, the feature dimensions map from $\mathbb{R}^{C\times H\times W}$ to $\mathbb{R}^{\text{hidden\_dim}\times H\times W}$, where hidden_dim is determined by the number of output channels in the pointwise convolution.

Wavelet transform effectively extracts multi-scale image details by decomposing the image into components of different frequencies. For a one-dimensional signal $f(t)$, its discrete wavelet transform (DWT) can be represented as:

$$\text{DWT}(f(t)) = \int_{-\infty}^{\infty} f(t)\psi_a(t)\,dt \tag{11}$$

where $\psi_a(t)$ is the wavelet basis function and a is the scaling parameter that controls the degree of frequency decomposition. In image processing, the two-dimensional discrete wavelet transform (DWT) decomposes the image into four subbands: low-frequency approximation coefficients (cA) and three high-frequency detail coefficients (cH, cV, and cD). This decomposition allows us to separately capture the global structure of the image (low-frequency part) and the fine details (high-frequency part). This multi-scale analysis enables the wavelet transform to efficiently capture image details in the frequency domain, especially in terms of edges, textures, and subtle changes.

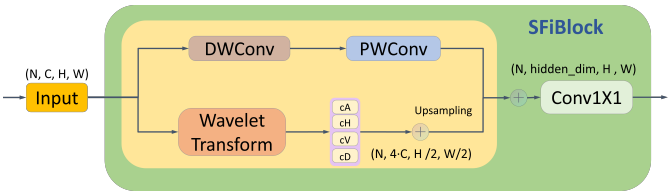

Figure 6: SFiBlock: Extracting Local Spatial Features with DSConv and Capturing Multi-Scale Frequency Domain Information with Wavelet Transform

The operation of VSSBlock module is described by the following equation:

$$y = x + \text{DropPath}(\text{SS2D}(\text{LN}(\mathcal{F}_{spatial} \oplus \mathcal{F}_{wavelet}))) \tag{12}$$

3.4. PCVSSBlock

In the **PCVSSBlock** module, the VSSBlock and WideFusion components are employed to capture both global and local information from the input features. The VSSBlock module enhances the network's global perception by combining local features with global context, achieved through modeling long-range dependencies. This approach enables the network to capture long-term relationships across the entire image. On the other hand, the WideFusion module specializes in extracting broadly focused features by mapping local features into a wider contextual space, which allows it to effectively capture fine-grained details from various receptive fields. This approach significantly strengthens the model's overall macro perception ability, especially when handling large-scale image contexts.

Subsequently a 1×1 convolution layer is then applied to fuse these features. This process effectively combines the complementary features extracted by both modules.

$$\text{Output} = \text{Conv}_{1\times1}(\text{VSSBlock}(x) \oplus \text{WideFusion}(x)) \quad (13)$$

By effectively merging features from different modules, the model can process multi-scale information simultaneously, which is particularly important for handling input images with varying resolutions. This fusion significantly enhances the model's overall performance in segmentation tasks.

3.5. Metric-Adaptive Loss and Dynamic Weight Adjustment

In medical image segmentation tasks, models often exhibit a tendency to prioritize larger, easily segmented regions while neglecting smaller and more intricate structures. For instance, in the ACDC cardiac segmentation dataset, the left ventricle (Lv), being relatively large and having strong contrast against the background, is segmented with high accuracy. However, this strong performance does not generalize well to more complex regions. For example, smaller structures like the myocardium (Myo) present fine and intricate details that models often fail to capture, resulting in lower segmentation accuracy for these regions.

To address this issue, we propose a novel **Metric-Adaptive Loss (MAL)** leveraging gated updates. MAL combines two critical metrics and a difficulty measure to dynamically balance the model's attention across different semantic regions. This design ensures the model allocates sufficient focus to small but crucial areas without compromising performance on larger regions, significantly enhancing segmentation quality in fine-grained areas and improving overall accuracy.

The difficulty measure is central to MAL, combining **Focal Loss** and **Dice Loss** to automatically identify challenging samples and assign them higher weights. By prioritizing these difficult areas, the model is encouraged to refine its segmentation of complex regions. The difficulty weight is computed as follows:

$$\mathcal{W}_{\text{Difficulty}} = \frac{1}{1 + \gamma(\mathcal{L}_{Dice} + \mathcal{L}_{Focal} + \epsilon)} \quad (14)$$

where γ is a hyperparameter that adjusts the influence of difficult samples. This dynamic weighting mechanism helps guide the model towards more accurate segmentation, particularly for small or complex structures that are traditionally harder to identify.

3.5.1. Uncertainty Measurement (UM)

Uncertainty Measurement (UM), often quantified using entropy, provides a means to evaluate the confidence of a model's predictions. Entropy measures the unpredictability or uncertainty inherent in the predicted probability distribution. A model confident in its predictions will exhibit low entropy, indicating a certain outcome, whereas high entropy reflects greater uncertainty in the prediction.

The entropy is calculated as $H(p) = -\sum p(c) \log p(c)$, where $p(c)$ represents the predicted probability of class c. To avoid issues related to computing log(0), we clamp the probability to a small range $p(c) = \text{clamp}(p(c) \in (1.0)$.

3.5.2. Gradient Information (GI)

To further enhance the adaptability of the loss function, we incorporate gradient information from the input image. Gradients effectively measure the intensity of local changes in the image, particularly in regions where there are significant variations in edges and details. In medical image segmentation tasks, especially when dealing with complex organs or structures, gradients help the model better identify boundary regions and fine structural features.

In this study, we use the **Laplacian operator** to compute the second-order gradients of the image, effectively revealing the intensity of regional changes and highlighting edge information. Specifically, the mathematical expression of the Laplacian operator is as follows: $\nabla^2 I = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}$.

Where I represents the input image, and ∇^2 is the Laplacian operator, representing the second-order spatial derivatives of the image. By computing the second-order derivatives for each pixel in the image, it measures the rate of intensity change at each pixel. Edge regions in an image typically have larger gradient values, as pixel intensity changes most rapidly at boundaries. To implement this, we use a 3×3 convolution kernel to apply the Laplacian operator:

$$K_{\text{lap}} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (15)$$

3.5.3. Dynamic Weight Adjustment

Dynamic Weight Adjustment is designed to dynamically adjust weights based on metrics like entropy, gradients, and difficult samples. This approach aims to direct the model's focus toward challenging, uncertain, or critical regions during training, enhancing segmentation performance. The formula for adaptive weight calculation is as follows:

$$\mathcal{W}_{\text{UM}} = \frac{1}{1 + \alpha H(p) + \epsilon} \quad (16)$$

$$\mathcal{W}_{\text{GI}} = \frac{1}{1 + \beta |\nabla^2 I| + \epsilon} \quad (17)$$

In this formulation, α controls the contribution of entropy, while β governs the influence of gradient information. The entropy term, $H(p)$, quantifies the uncertainty in the model's prediction for each pixel. Higher entropy values indicate regions where the model is less confident, signaling the need for increased attention during training. Gradient information, $|\nabla^2 I|$, represents the magnitude of intensity variation in the image, often corresponding to edges and other significant structural features. The Laplacian operator (∇^2) captures this second-order derivative information, revealing areas of rapid intensity change, which are crucial for accurate segmentation. Regions with larger Laplacian values are typically more critical for accurate segmentation. Additionally, the difficulty measurement dynamically adjusts the loss weight for challenging samples, encouraging the model to prioritize hard-to-segment regions. This mechanism ensures that the model focuses on areas that are essential for improving overall segmentation performance.

The final adaptive weight for each sample is the product of the individual weights:

$$\mathcal{W} = \mathcal{W}_{\text{UM}} \cdot \mathcal{W}_{\text{GI}} \cdot \mathcal{W}_{\text{Difficulty}} \tag{18}$$

This product form ensures that the model focuses on the critical regions of the image during training, rather than treating all regions uniformly. It helps prevent the model from overemphasizing easily segmentable areas while neglecting more challenging, smaller regions.

To avoid excessive weight fluctuations that may lead to instability during training, we incorporate weight smoothing. By computing an exponentially weighted average between the current and previous weights, we smooth the weight updates, thereby enhancing the stability of the training process. The formula for the smoothed weights is as follows:

$$\mathcal{W}_{\text{final}} = \lambda \cdot \mathcal{W}_{\text{current}} + (1 - \lambda) \cdot \mathcal{W}_{\text{last}} \tag{19}$$

Here, λ is the smoothing factor, typically set to 0.1. The smoothing process effectively mitigates issues such as gradient explosion or vanishing gradients due to excessive weight fluctuations, ensuring better convergence and stability during training.

The final total loss is computed as the weighted sum of Focal Loss and Dice Loss, with adaptive weights applied. This loss function effectively amplifies the contribution from difficult, uncertain, or critical regions, guiding the model to perform more accurate segmentation in these areas. The total loss is calculated as:

$$\mathcal{L}_{MAL} = \mathcal{W} \cdot (\mathcal{L}_{Dice} + \mathcal{L}_{Focal}) \tag{20}$$

This weighted loss function enables the model to effectively prioritize fine-grained and challenging regions during training, thereby improving overall segmentation performance. It enhances segmentation accuracy and robustness, particularly for small and complex structures in medical images.

4. Experiment

4.1. Implementation Details

We conducted our experiments using two NVIDIA Titan V and Titan XP GPUs, both equipped with 12GB of memory each, using CUDA version 11.8. The model training was implemented in PyTorch 2.0.0, employing the Distributed Data Parallel (DDP) strategy for efficient GPU utilization. Input images were resized to 224×224 pixels and standardized to promote stable training. The dataset was divided into training and validation sets in an 80:20 ratio. The AdamW optimizer was used, incorporating weight decay to enhance generalization. A cosine annealing learning rate schedule was applied, starting with an initial learning rate of 3e-4. To further improve model robustness and generalization, data augmentation techniques, such as affine transformations, image flipping, and rotation, were employed during training.

4.2. Datasets and Evaluation Metrics

To comprehensively evaluate the performance of our proposed model, we conducted experiments across four representative medical imaging modalities and nine associated datasets, including Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Ultrasound Imaging, and Dermatoscopic Imaging. These modalities cover a wide range of common medical image types, enabling us to assess the generalization capability of our model in multi-modal environments.

To comprehensively assess the performance of our model in medical image segmentation tasks, we utilized several standard evaluation metrics, including the Dice coefficient, Intersection over Union (IoU), Accuracy (ACC), Specificity (SP), and Sensitivity (SE). These metrics provide a multidimensional evaluation of the segmentation results, capturing precision, robustness, and comprehensiveness, and ensuring that the overall performance of our model is effectively represented across various datasets.

Acknowledging that different models may exhibit strengths in specific organ segmentation tasks, we included a range of state-of-the-art (SOTA) models in our comparative experiments. These models comprise Transformer-based architectures, U-Net variants, and various adaptations of the Mamba architecture, representing the leading methodologies in the field. This selection ensures a comprehensive evaluation across a diverse spectrum of model capabilities. Through comparisons with these advanced models, we demonstrate the superior performance of our proposed Mamba paradigm in multi-modal segmentation tasks.

4.2.1. ACDC Dataset

The ACDC (Automated Cardiac Diagnosis Challenge) dataset contains cardiac MRI scans from 100 patients, divided into five subgroups (Normal, MINF, DCM, HCM, ARV), with 30 cases per group. Each case includes a 4D MRI of a full cardiac cycle with detailed annotations at the end-diastolic (ED) and end-systolic (ES) frames. The subtle boundaries between the left ventricle, right ventricle, and myocardium present significant challenges for accurate segmentation tasks.

4.2.2. MM-WHS2017-CT Dataset

The MM-WHS2017 dataset[24] contains 120 multimodal cardiac images (60 CT/CTA and 60 MRI) from the MICCAI 2017 Challenge, covering seven cardiac structures: LV, LA, RV, RA, MYO, AA, and PA. The data are anonymized and ethically approved. Cross-sectional, coronal, and sagittal slices were extracted for evaluation, with the cross-sectional view as the primary focus.

4.2.3. LAHeart2018 Dataset

The LAHeart2018 dataset[25] contains 154 3D MRI images of atrial fibrillation patients, with a grayscale range of [0, 255] and a spatial resolution of 0.625×0.625×0.625 mm³ across 88 Z-axis slices. Although the left atrium segmentation is a monosemantic task, accurately segmenting its fine structures remains challenging.

4.2.4. Synapse Dataset

The Synapse dataset[26] consists of 3,779 high-quality abdominal axial CT images from 30 patients. It provides comprehensive coverage of key abdominal organs, including the aorta, gallbladder, spleen, kidneys, liver, pancreas, and stomach, serving as a valuable resource for automatic organ segmentation and diagnostic research.

4.2.5. ISIC Dataset

The ISIC2016[27] and ISIC2017[28] datasets, released by the International Skin Imaging Collaboration (ISIC), support melanoma diagnosis research. ISIC2016 contains 1,279 images with a color depth of 24-bit and dimensions ranging from 722×542 to 4,288×2,848 pixels. The larger ISIC2017 dataset includes 2,750 images, with sizes varying from 540×576 to 4,499×6,748 pixels, and all images are fully annotated.

4.2.6. PH2 Dataset

The PH2 dataset[29] consists of 200 high-quality dermoscopic images of skin lesions, sourced from the Pedro Hispano Hospital in Matosinhos, Portugal. It features common lesion types like melanocytic nevi, melanoma, and basal cell carcinoma, with annotations by experienced dermatologists to ensure accuracy. The fuzzy and complex boundaries pose challenges for precise segmentation.

4.2.7. DDTI Dataset

The DDTI[30] dataset consists of 637 ultrasound thyroid images with pixel-level annotations, provided by Pedraza et al. This dataset holds significant value in the analysis of thyroid ultrasound images, encompassing a diverse range of cases including thyroiditis, goiter, nodules, and cancer. The complexity of these images reflects the challenges of nodule detection.

4.2.8. TN3K Dataset

TN3K[31] is a challenging dataset for thyroid nodule segmentation, comprising 3,493 ultrasound images with precise pixel-level nodule mask annotations. The images are sourced from various devices and multiple views, ensuring diversity and quality in the dataset. This dataset was collected by Zhujiang Hospital, Southern Medical University.

4.3. Experimental Results

4.3.1. Results on ACDC Dataset

On the ACDC cardiac segmentation dataset, the DiscWideFusion module achieves a computational complexity of 30.6 GFlops and a parameter count of 21.4M, demonstrating performance competitive with current state-of-the-art models in cardiac segmentation. The RepLKWideFusion module achieved an average Dice coefficient of 93.77%, significantly outperforming all other models, particularly in segmenting the left ventricle (96.61%) and right ventricle (94.43%), demonstrating exceptional accuracy. Compared to recent models based on CNN, Transformer, and Mamba architectures (e.g., LHU-Net (92.66%), FCT (93.02%), and UU-Mamba (92.79%)), our model sets a new SOTA record for multi-semantic region segmentation of the heart, surpassing the previous performance of FCT. Additionally, we present a comparative analysis of the model's segmentation performance (Dice coefficient) in relation to the number of parameters and computational complexity (GFlops); the corresponding results are shown in Figure 2.

Table 1: Performance Comparison of Different Methods on the ACDC Dataset (**Bold**: Best, Underline: Second Best)

Methods	Year	Avg	RV	Myo	LV
SwinUNet[12]	2021	90.00	88.55	85.62	95.83
MISSFormer[32]	2021	90.86	89.55	88.04	94.99
nnUNet[6]	2018	91.61	90.24	89.24	95.36
TransCASCADE	2023	91.63	89.14	90.25	95.50
Mamba-Unet[33]	2024	91.08	90.80	88.09	94.45
MS-TCNet[34]	2024	91.43	89.43	89.09	95.77
PVT-CASCADE[35]	2023	91.46	88.90	89.97	95.50
TC-CoNet[36]	2023	91.58	90.27	88.98	95.47
UD-Mamba[37]	2024	91.99	90.85	90.69	94.45
LUCF-Net[38]	2024	92.19	90.46	90.13	95.96
U-Mamba	2024	92.22	91.83	90.22	94.54
Parallel MERIT[39]	2023	92.32	90.87	90.00	96.08
LHU-Net[40]	2024	92.66	91.15	90.56	96.26
UU-Mamba[41]	2024	92.79	92.41	**90.90**	95.04
BATFormer[42]	2022	92.84	91.97	90.26	96.30
FCT[43]	2023	93.02	92.64	90.51	95.90
Ours_Disc	2024	93.59	94.12	90.32	96.33
Ours_Rep	2024	**93.77**	**94.43**	90.27	**96.61**

4.3.2. Results on MM-WHS2017 Dataset

In experiments conducted on the MM-WHS 2017 dataset, we evaluated state-of-the-art models, including Transformer-based and traditional convolutional networks. While Transformer architectures have made significant strides in image segmentation tasks, traditional convolutional networks still perform exceptionally, especially in managing complex boundaries and blurry regions due to the unique properties of convolution operations. The comparison between U-Net and Transformer architectures highlights convolutional networks' effectiveness in cardiac and vascular segmentation.

The DiscWideFusion model achieved a Dice coefficient of **89.65%** for the Myo region, outperforming all other models and demonstrating its accuracy in segmenting cardiac regions. The RepLKWideFusion model achieved a Dice coefficient of **97.53%** for the AA region, confirming its superiority in complex region segmentation. The RepLKWideFusion model achieved an average Dice coefficient of **94.04%**, surpassing existing models and showcasing the benefits of the hybrid convolutional and Mamba architecture in multi-semantic segmentation. These results show that our model excels in multi-region cardiac and vascular segmentation, especially in small and com-

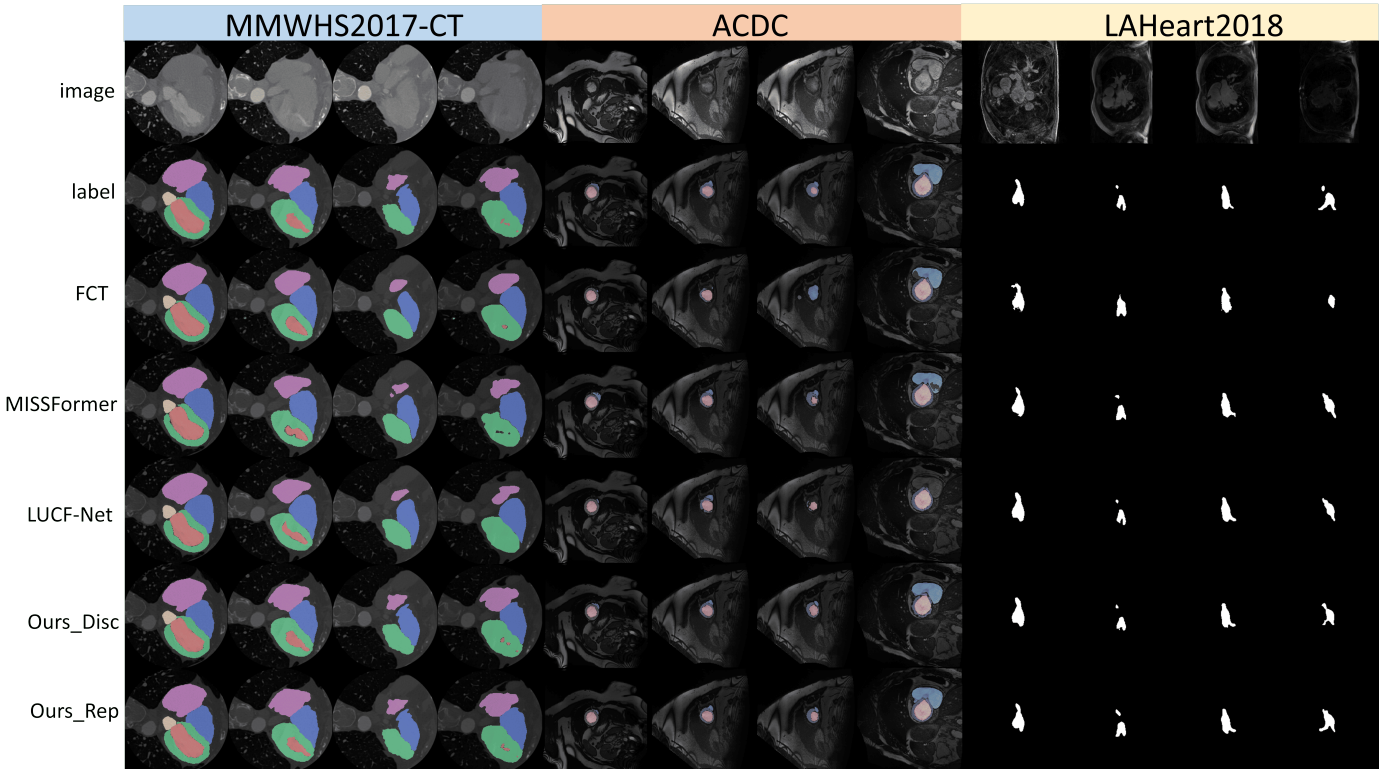

Figure 7: Experiments conducted on the MMWHS-CT, ACDC, and LAHeart2018 cardiac datasets demonstrate the model's effectiveness in handling ambiguous boundaries and fine structural details in both single-region and multi-region segmentation tasks, as well as in single- and multi-class segmentation across diverse cardiac imaging modalities, including MRI and CT.

Table 2: Performance Comparison of Different Methods on the MMWHS2017-CT dataset (**Bold**: Best, Underline: Second Best)

Methods	Year	Myo	LA	LV	RA	RV	AA	PA	Avg
Swin-Unet	2021	68.90	84.15	82.07	74.87	72.73	91.02	76.08	78.55
PVT-V2-b3[44]	2021	76.46	82.14	86.97	72.24	74.89	85.85	75.51	79.15
V-net[45]	2016	79.10	85.30	81.30	90.90	81.60	76.30	71.70	80.89
KTH	2017	85.60	93.00	92.30	87.10	85.70	89.4	83.50	88.09
MISSFormer	2021	85.63	93.38	92.48	84.79	84.68	96.35	84.19	88.79
Unet	2015	86.86	92.93	94.03	86.29	87.96	95.87	82.56	89.50
GUT	2017	88.10	92.90	91.80	88.80	90.90	93.30	84.00	89.97
FCT	2023	88.77	94.30	92.06	87.91	**92.19**	93.49	81.68	90.06
LUCF-Net	2024	88.80	93.64	93.16	86.58	90.74	96.44	83.79	90.44
MAUNet[46]	2020	89.30	91.00	92.50	92.80	88.60	92.50	86.60	90.47
Ours_Disc	2024	**89.65**	94.97	94.43	91.07	91.68	97.20	90.99	92.86
Ours_Rep	2024	89.57	**96.19**	**95.16**	**95.22**	92.14	**97.53**	**92.47**	**94.04**

plex areas.

4.3.3. Results on LAHeart2018 Dataset

In the experiments on the LAHeart2018 dataset, we evaluated multiple models, and our DiscWideFusion and RepLKWideFusion versions demonstrated superior segmentation performance. The DiscWideFusion version achieved a Dice coefficient of 93.85% and an IoU of 88.44%, outperforming other models, especially in segmenting fine left atrium structures. The fine structures in the left atrium (such as the atrial appendages and septum) pose challenges due to complex morphology and blurred boundaries. Despite a slightly lower Dice score, RepLKWideFusion achieved an IoU of 87.98%, maintaining high precision and showcasing its advantage in small region segmentation.

Table 3: Performance Comparison of Different Methods on the LAHeart2018 dataset (**Bold**: Best, Underline: Second Best)

Methods	Year	Dice	IOU
Unet	2015	85.15	74.24
LUCF-Net	2024	89.52	81.11
MISSFormer	2021	89.96	81.80
FAT-Net[47]	2022	90.42	82.54
Swin-Unet	2021	90.57	87.81
Ours_Disc	2024	**93.85**	**88.44**
Ours_Rep	2024	93.32	87.98

Our models outperformed those submitted to the 2018 Atria Segmentation Challenge, where the highest Dice coefficient was 93.20%. They excelled in segmenting small, mono-semantic regions of the left atrium, particularly in difficult-to-segment fine structures. This superior performance is attributed to the multi-scale feature extraction and fusion capabilities of the WideFusion module, which significantly enhanced the segmentation accuracy of fine structures. These results highlight our method's advantage in segmenting small structures and complex anatomical regions, particularly those with intricate morphology and unclear boundaries.

4.3.4. Results on TN3K and DDTI datasets

Ultrasound image segmentation faces challenges from high noise and blurred boundaries, especially with fine structures and complex backgrounds. Traditional models often perform poorly under such conditions. Our proposed models, DiscWideFusion and RepLKWideFusion, effectively address these issues. Using the Mamba framework's long-sequence modeling and wavelet transforms for high-frequency extraction, our models excel in capturing fine details and accurately segmenting target regions in noisy, blurred environments.

On the TN3K dataset, DiscWideFusion achieved a Dice score of **89.29%** and an IoU of 70.67%, while RepLKWideFusion reached an IoU of 74.95%, showing strong performance in complex ultrasound images. Compared to traditional models like UNet and Swin-Unet, our approach outperforms in noisy and unclear boundary regions, especially for small lesion segmentation.

Wavelet transforms enhance robustness by extracting high-frequency details for precise edge detection and better understanding of complex structures. The Mamba framework's long-sequence modeling strengthens feature extraction, enabling accurate segmentation of diverse structures.

These results show that our models, using the Mamba framework and wavelet transforms, provide significant advantages in ultrasound image segmentation, particularly under high noise and blurred boundary conditions, enhancing precision for fine structures.

4.3.5. Results on Synapse datasets

Experimental results on the Synapse multi-organ dataset show that our models excel in segmenting multiple organs, es-

Table 4: Performance Comparison of Different Methods on the Thyroid Ultrasound Image Dataset (**Bold**: Best, Underline: Second Best)

Methods	Year	TN3K		DDTO	
		Dice	**IOU**	**Dice**	**IOU**
SwinUnet	2021	46.02	38.18	46.70	30.36
UNet	2015	79.51	65.99	59.74	42.59
SGUNet[48]	2021	79.55	66.05	69.92	45.90
SegNet[49]	2015	79.91	66.54	65.19	48.36
FCN[50]	2015	81.08	68.18	69.96	53.80
TRFE[51]	2023	81.19	68.33	69.04	52.72
TransUNet[11]	2021	81.84	69.26	74.43	59.28
CPFNet[52]	2020	82.70	70.50	74.77	59.70
Deeplabv3+[53]	2018	82.77	70.60	74.40	59.23
TRFE+	2023	83.30	71.38	75.37	60.47
LUCF-Net	2024	84.86	69.89	75.83	58.49
FAT-Net	2024	85.18	65.80	75.19	59.86
Ours_Disc	2024	**89.29**	70.67	77.33	62.97
Ours_Rep	2024	87.36	**74.95**	**77.43**	**63.79**

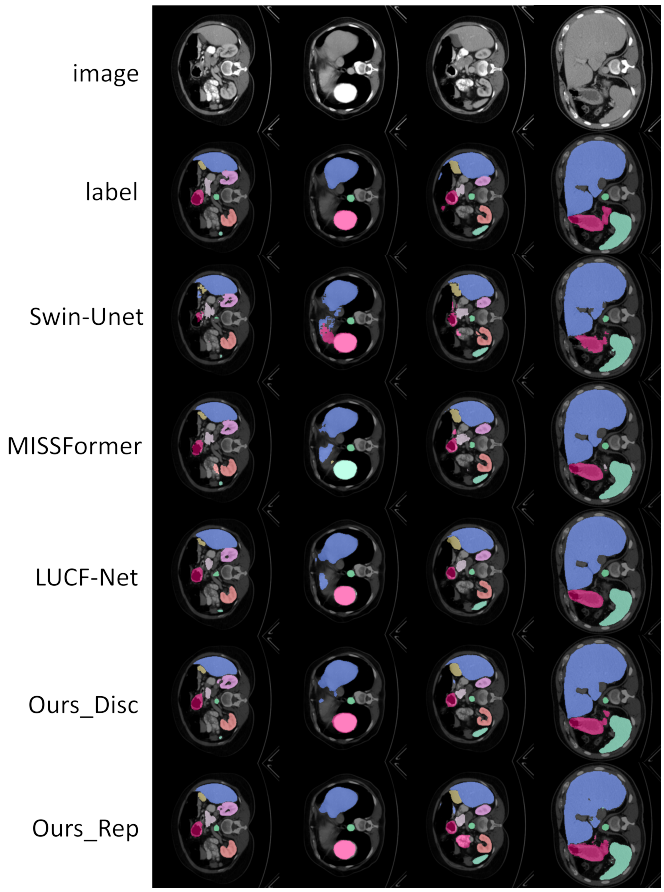

Figure 8: Segmentation results of various methods on the Synapse multi-organ CT dataset, illustrating our models' ability to accurately delineate complex structures and achieve precise results, even in challenging backgrounds.

Table 5: Performance Comparisons Of Various Techniques On Synapse Abdominal Multi-Organ Dataset. (**Bold**: Best, Underline: Second Best)

Methods	Year	Aorta	Gallbladder	Kidney(R)	Kidney(L)	Liver	Pancreas	Spleen	Stomach	Avg
UNet	2015	89.07	69.72	68.60	77.77	93.43	53.98	86.67	56.98	76.85
TransUnet	2021	87.23	63.13	77.02	81.87	94.08	55.86	85.08	75.62	77.48
Swin-Unet	2021	85.47	66.53	79.61	83.28	94.29	56.58	90.66	76.60	79.13
PVT-CASCADE	2023	83.01	70.59	80.37	82.23	94.08	64.43	90.01	83.69	81.06
VM-UNet	2024	86.40	69.41	86.16	82.76	94.17	58.80	89.51	81.40	81.08
MISSFormer	2021	86.99	68.65	82.00	85.21	94.41	65.67	91.92	80.81	81.96
SegFormer3D[54]	2024	90.43	55.26	86.13	86.53	95.68	73.06	89.02	81.12	82.15
LUCF-Net	2024	89.66	71.14	85.04	87.88	95.52	67.32	91.54	85.05	84.22
MERIT[39]	2022	87.71	74.40	87.79	84.85	95.26	71.81	92.01	85.38	84.90
nnFormer[55]	2021	**92.04**	70.17	86.57	86.25	**96.84**	**83.35**	90.51	86.83	86.57
Ours_Disc	2024	89.35	75.77	92.57	93.50	95.49	73.17	92.07	**89.99**	87.74
Ours_Rep	2024	90.14	**77.16**	**94.24**	**94.68**	96.59	75.00	**95.32**	81.78	**88.11**

pecially complex anatomical structures. The RepLKWideFusion model achieves an average Dice score of 88.11%, with outstanding right kidney (94.68%) and spleen (95.32%) segmentation results. These results highlight the model's ability to accurately delineate intricate structures and maintain high precision, even in challenging backgrounds.

The DiscWideFusion model, with an average Dice score of 87.74%, shows strong segmentation performance for certain organs. The liver segmentation achieves a Dice score of 95.49%, showcasing the model's precision in segmenting large organs. The gallbladder and stomach segmentations achieve Dice scores of 75.77% and 89.99%, respectively, showing competitive performance. Our approach outperforms other Transformer and Mamba-based models in various segmentation tasks.

Table 6: Performance Comparison of Different Methods on the ISIC2016 dataset (**Bold**: Best, Underline: Second Best)

Methods	Year	Dice	IOU	ACC	SE	SP
Unet	2015	88.57	81.53	94.37	91.76	95.71
Attention-Unet[56]	2018	88.75	81.58	94.14	90.31	96.45
UNext[57]	2022	90.48	83.66	-	-	-
BAT	2021	90.63	86.76	94.89	91.25	95.69
DuAT[58]	2022	90.65	86.89	95.09	97.99	94.88
DAGAN[59]	2022	90.85	84.42	95.82	92.28	95.68
SLP-Net[60]	2024	91.12	-	95.70	91.45	96.94
LUCF-Net	2024	91.18	84.98	96.18	93.38	97.75
FAT-Net	2022	91.59	85.30	96.04	92.59	96.02
GFANet[61]	2023	91.78	85.92	96.04	92.95	97.25
EIU-Net	2023	91.90	85.50	95.90	91.80	94.30
Ms RED[62]	2022	92.66	87.03	96.42	-	-
MobileUNETR[63]	2024	92.80	87.47	96.59	93.03	96.87
TC-Net[64]	2023	92.82	86.68	96.06	93.17	97.12
Ours_Disc	2024	93.22	87.98	96.28	92.72	96.53
Ours_Rep	2024	**93.40**	**88.16**	**96.97**	**93.10**	**98.03**

4.3.6. Results on Skin Lesion Image datasets

Dermatological datasets face challenges including diverse lesion areas, inconsistent image quality, and blurred boundaries. Although many models can accurately locate lesion areas, they often struggle with fine boundary delineation. Our model demonstrates significant advantages across datasets, especially in detail segmentation and handling complex boundaries.

On the ISIC 2016 and ISIC 2017 datasets, RepLKWideFusion achieved average Dice scores of 93.40% and 90.81%, outperforming state-of-the-art methods, especially in complex lesion boundary handling. The model also excelled in accuracy (ACC) and specificity (SP), reaching 96.97% and 98.03% on ISIC2016, indicating its strong ability to identify lesions and differentiate normal skin. On the PH2 dataset, the model achieved a Dice score of 96.37%, IoU of 93.04%, and ACC of 98.71%, further validating its fine lesion segmentation and normal skin differentiation. The model also achieved a sensitivity (SE) of 98.75%, demonstrating its effectiveness in detecting lesions and minimizing false negatives.

In summary, our model delivers high precision and robust segmentation across dermatological datasets, excelling in fine-grained tasks and showing significant practical potential.

4.3.7. FeatureMap Analysis

To evaluate the effectiveness of the WideFusion module, VSS_Block, and their fused representation, we conducted a detailed analysis of feature maps at various network stages, including Encoder1–4, Decoder1–4, and the BottleNeck. Specifically, we visualized the feature maps generated by **WideFusion**, **VSSBlock**, and their fused representation (**PCVSSBlock**) to explore their individual contributions to the segmentation task.

The analysis shows that **WideFusion** focuses on capturing multi-scale local structures, such as subtle edges and boundaries, essential for fine-grained segmentation. VSSBlock, on the other hand, excels at modeling global semantic information and contextual relationships by using wavelet transformations,

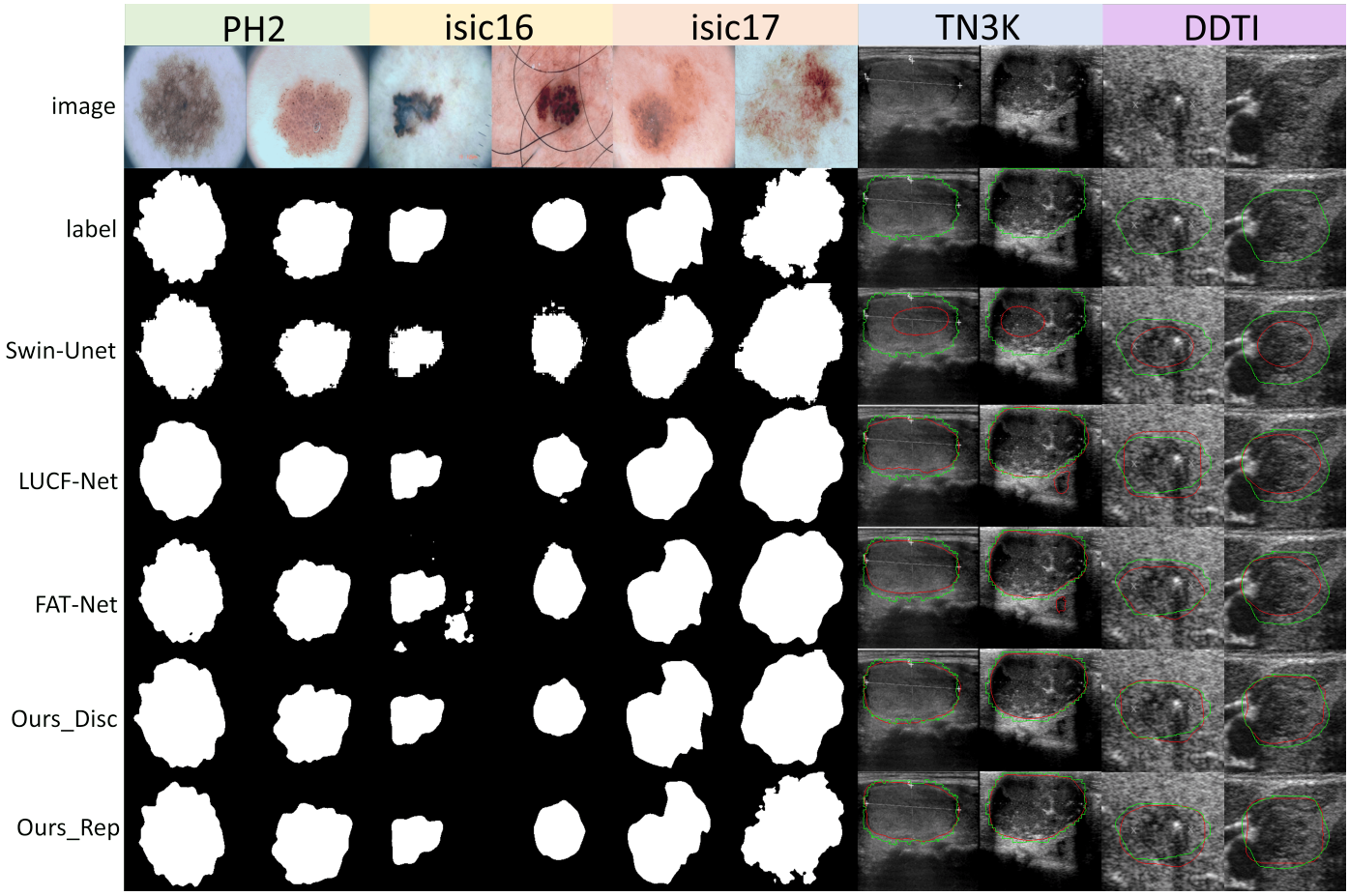

Figure 9: Results from the Skin Lesion and Thyroid Ultrasound Image datasets. Despite challenges such as diverse, blurred lesion areas and the high noise in ultrasound images, our models achieve precise segmentation.

Table 7: Performance Comparison of Different Methods on the ISIC2017 dataset (**Bold**: Best, Underline: Second Best)

Methods	Year	Dice	IOU	ACC	SE	SP
UNet	2015	81.59	72.34	91.64	81.72	96.80
Attention-Unet	2018	80.82	71.73	91.45	79.98	97.76
DAGAN	2022	84.25	75.94	93.04	83.63	97.16
GFANet	2023	85.74	77.75	93.97	81.37	97.87
FAT-Net	2022	85.00	76.53	93.26	83.92	97.25
Ms RED	2022	86.48	78.55	94.10	-	-
MobileUNETR	2024	86.84	79.00	94.46	85.18	96.93
EGE-UNet[65]	2023	88.77	79.81	96.42	89.31	98.16
LeaNe[66]	2024	88.89	78.93	95.72	90.63	97.72
VM-Unet-2[67]	2024	90.45	82.34	87.68	87.69	98.49
VM-Unet	2024	90.70	80.23	96.45	88.37	98.42
LightM-Unet[68]	2024	90.80	-	95.83	88.39	98.46
LUCF-Net	2024	91.16	83.79	92.74	93.23	97.77
MUCM-Net[69]	2024	**91.20**	-	96.72	88.24	98.29
Ours_Disc	2024	90.15	82.75	94.86	93.98	98.72
Ours_Rep	2024	90.81	**84.10**	**96.95**	**94.72**	**99.48**

Table 8: Performance Comparison of Different Methods on the PH2 dataset (**Bold**: Best, Underline: Second Best)

Methods	Year	Dice	IOU	ACC	SE	SP
Unet	2015	89.00	81.70	93.16	90.66	95.07
Attention-Unet	2018	90.03	85.82	92.76	92.05	96.40
DAGAN	2022	92.01	-	94.25	83.20	96.40
HiFormer[70]	2023	94.27	89.48	95.49	94.98	94.18
FAT-Net	2022	94.40	89.62	97.03	94.41	97.41
Ms RED	2022	94.65	90.14	96.80	-	-
MALUNet[71]	2022	94.76	90.06	-	94.23	97.72
AttSwinUNet[72]	2022	95.04	-	96.85	94.39	95.76
GFANet	2023	95.06	90.98	97.09	96.08	97.57
UNext	2022	95.16	90.81	-	95.24	97.75
ULFAC-Net[73]	2023	95.29	91.28	97.01	95.63	97.42
TransCS-Net[74]	2023	95.30	91.09	96.74	95.53	96.58
LUCF-Net	2024	95.50	91.44	97.25	96.51	95.81
MobileUNETR	2024	95.70	92.30	97.71	96.05	96.60
Ours_Disc	2024	95.83	92.51	97.98	97.63	97.36
Ours_Rep	2024	**96.37**	**93.04**	**98.71**	**98.75**	**98.91**

enabling a comprehensive understanding of overall structure and key boundaries. The fused PCVSSBlock feature map combines the strengths of both modules, preserving local detail sensitivity while incorporating global context. This complementary effect underscores the benefits of the two modules working together, striking a balance between local precision and global consistency in segmentation.

We also analyzed the multi-scale input-output characteristics of our U-Net-based model. We used upsampling to align the feature maps from various decoder depths with the original image size, generating the DS-Out 1–3 outputs. Although the understanding and recovery abilities of the decoders differ by level, all decoders demonstrate strong recovery performance at their respective feature levels. This indicates that the U-Net architecture has strong multi-scale reconstruction capabilities across different decoder levels, supporting segmentation tasks effectively.

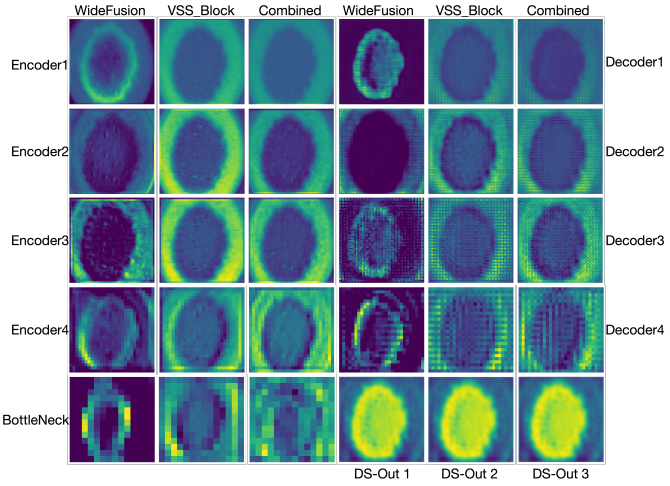

Figure 10: Comparison of Local and Global Feature Map Representations: WideFusion, VSSBlock, Fused Representations, and DS-Out 1–3 Outputs

4.4. Ablation Study

4.4.1. Ablation Experiment on WideFusion Module

Table 9: Ablation Study of Convolution Selection in RepLKWideFusion Module

Methods	RV	Myo	LV	Avg	GFlops	Params(M)
[3×3, 5×5, 7×7]	94.12	**90.32**	96.33	93.59	**30.6**	**21.4**
[5×5, 7×7, 9×9]	91.15	87.20	95.50	91.02	**30.6**	**21.4**
[3×3, (5, 3), (7, 3)]	**94.43**	90.27	**96.61**	**93.77**	68.8	47.2
[3×3, (7, 3), (9, 3)]	92.49	89.68	95.85	92.50	97.7	66.7
[(5, 3), (7, 3), (9, 3)]	92.43	90.11	96.14	92.55	110.6	75.4

This ablation study investigates the impact of various convolutional kernel configurations in the WideFusion module, comparing traditional discrete convolutions in DiscWideFusion with the reparameterized kernel design in RepLKWideFusion.

Using a combination of three discrete convolutional kernels ([5×5, 7×7, 9×9]), the model achieved an average DSC of 93.59%. This configuration effectively captures local details and maintains low computational complexity but struggles with fine detail capture in complex images. In contrast, the reparameterized kernel configuration, such as [3×3, (5, 3), (7, 3)], showed significant advantages. In this design, (5, 3) represents a combination of a 5×5 large kernel and a 3×3 small kernel. Combining a larger kernel (5×5) with a smaller one (3×3) expands the receptive field while maintaining low computational complexity and capturing richer contextual information. The reparameterization design leverages the complementary strengths of large and small kernels, enhancing the model's ability to capture fine details and model global contexts. This configuration achieved an average DSC of 93.77%, with segmentation accuracies for the right ventricle (RV) and left ventricle (LV) at 94.43% and 96.61%, respectively.

Small kernels, such as 3×3, are effective at capturing fine details and delineating boundaries, while larger kernels, such as 7×3, increase the receptive field, allowing for a broader contextual understanding. This strategic combination enhances the model's ability to perform precise segmentation in complex medical image regions. Although the reparameterized configuration introduces an increase in computational complexity—evidenced by the rise in GFlops and the number of parameters to 68.8 and 47.2M, respectively—the performance improvement, especially in fine-grained segmentation tasks, is considerable.

4.4.2. Ablation Experiment on Loss Function Module

This experiment evaluates the generalizability of the proposed Metric-Adaptive Loss (MAL) by testing it on our model, the baseline U-Net, and FAT-Net, which is one of the top-performing models in the current batch. The experiment uses a skin lesion segmentation dataset containing lesions with complex boundaries and diverse morphologies, making segmentation more challenging. The results show that MAL effectively improves segmentation performance across various model architectures, confirming its strong generalizability.

Table 10: Ablation Study of Loss Function

Methods	Loss	ISIC2016		ISIC2017		PH2	
		Dice	IOU	Dice	IOU	Dice	IOU
Unet	MAL	88.57	81.53	83.96	72.34	89.00	81.70
	Dice Loss	86.52	75.13	79.98	63.43	88.58	79.50
	Focal Loss	87.49	77.83	81.59	72.22	87.55	76.13
FAT-Net	MAL	92.03	85.98	86.18	76.75	94.95	90.38
	Dice Loss	91.59	85.30	85.00	76.53	94.50	89.57
	Focal Loss	91.39	84.21	83.35	74.07	94.40	89.62
Ours_Disc	MAL	93.22	87.98	90.15	82.75	95.83	92.51
	Dice Loss	92.96	86.89	86.18	76.75	94.40	89.55
	Focal Loss	91.92	85.47	85.40	75.08	93.19	87.39
Ours_Rep	MAL	93.40	88.16	90.81	84.10	96.37	93.04
	Dice Loss	92.90	86.76	86.08	75.35	91.68	84.96
	Focal Loss	90.37	80.02	90.32	83.73	93.95	88.77

5. Conclusion

The **PCMamba** architecture introduced in this study integrates the Mamba model with optimized convolutional structures, significantly boosting both coarse and fine-grained performance in medical image segmentation tasks. For the Mamba model, we designed **SFiBlock**, which enables joint feature extraction in both time and frequency domains using DSConv and wavelet transforms, strengthening Mamba's ability to model long-range dependencies. Additionally, we introduced two innovative modules for the convolutional structure: the **DiscWideFusion** module, which relies on discrete convolutions and features a smaller number of parameters and reduced computational complexity; and the **RepLKWideFusion** module, which incorporates large kernel convolution reparameterization techniques, focusing on improving segmentation accuracy and accelerating large kernel convolution through FFTConv. Moreover, we proposed the Metric-Adaptive Loss (MAL) function, which dynamically adapts training parameters based on key metrics, significantly enhancing segmentation accuracy, particularly for complex regions like low-contrast boundaries.

Extensive validation on multiple medical datasets demonstrates that PCMamba achieves exceptional efficiency and strong generalization ability. The model has achieved SOTA performance across multiple tasks, proving its stability and outstanding performance in medical image processing. Our research explores an innovative method that parallelizes the Mamba architecture with convolutional operations, achieving an effective trade-off between parameter size, computational complexity, and segmentation performance. This approach successfully addresses the challenge of balancing coarse and fine-grained segmentation in medical image tasks and provides a new perspective for future work, while offering a robust baseline for related research.

In future work, we aim to delve deeper into dynamic kernel adaptation mechanisms to further enhance PCMamba's flexibility and adaptability. Specifically, we will dynamically adjust the size and shape of the convolutional kernels based on local information in the input feature maps, enabling the model to adaptively select the most suitable convolutional operation for the current image characteristics. Furthermore, the scanning mechanism of Mamba will also be explored as a potential innovation direction in future research.

6. ACKNOWLEDGMENTS

This work has been supported by the Natural Science Foundation of Shandong Province (No. ZR2019MF011). This work was also supported by the National Natural Science Foundation of China (No. 62376136 and No. 62076149), Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515011935) . We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V used for this research.

7. CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, P.-M. Jodoin, Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?, IEEE Transactions on Medical Imaging 37 (11) (2018) 2514–2525. doi:10.1109/TMI.2018.2837502.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, https://arxiv.org/abs/1512.03385v1 (2015).

[3] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, https://arxiv.org/abs/2201.03545v2 (2022).

[4] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, S. Xie, ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders, https://arxiv.org/abs/2301.00808v1 (2023).

[5] Q. Jin, Z. Meng, C. Sun, L. Wei, R. Su, RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans, https://arxiv.org/abs/1811.01328v1 (2018). doi:10.3389/fbioe.2020.605132.

[6] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, K. H. Maier-Hein, nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation, https://arxiv.org/abs/1809.10486v1 (2018).

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, https://arxiv.org/abs/2010.11929v2 (2020).

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, https://arxiv.org/abs/2103.14030v2 (2021).

[9] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions, https://arxiv.org/abs/2102.12122v2 (2021).

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, https://arxiv.org/abs/1409.0575v3 (2014).

[11] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, https://arxiv.org/abs/2102.04306v1 (2021).

[12] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation (2021). arXiv:2105.05537.

[13] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, V. M. Patel, Medical Transformer: Gated Axial-Attention for Medical Image Segmentation, https://arxiv.org/abs/2102.10662v2 (2021).

[14] Z. Yu, S. Han, Z. Song, 3D Medical Image Segmentation based on multi-scale MPU-Net, https://arxiv.org/abs/2307.05799v2 (2023).

[15] A. Gu, K. Goel, C. Ré, Efficiently Modeling Long Sequences with Structured State Spaces, https://arxiv.org/abs/2111.00396v3 (2021).

[16] A. Gu, T. Dao, Mamba: Linear-Time Sequence Modeling with Selective State Spaces, https://arxiv.org/abs/2312.00752v2 (2023).

[17] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, Y. Liu, VMamba: Visual State Space Model (2024). arXiv:2401.10166, doi:10.48550/arXiv.2401.10166.

[18] J. Ma, F. Li, B. Wang, U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation, https://arxiv.org/abs/2401.04722v1 (2024).

[19] J. Ruan, S. Xiang, VM-UNet: Vision Mamba UNet for Medical Image Segmentation, https://arxiv.org/abs/2402.02491v1 (2024).
[20] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, Y. Yu, Y. Liang, G. Shi, S. Zhang, H. Zheng, S. Wang, Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining, https://arxiv.org/abs/2402.03302v2 (2024).
[21] Large Window-based Mamba UNet for Medical Image Segmentation: Beyond Convolution and Self-attention, https://arxiv.org/html/2403.07332v1.
[22] Z. Wang, C. Ma, Weak-Mamba-UNet: Visual Mamba Makes CNN and ViT Work Better for Scribble-based Medical Image Segmentation, https://arxiv.org/abs/2402.10887v1 (2024).
[23] X. Ding, X. Zhang, Y. Zhou, J. Han, G. Ding, J. Sun, Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs, https://arxiv.org/abs/2203.06717v4 (2022).
[24] X. Zhuang, L. Li, C. Payer, D. Štern, M. Urschler, M. P. Heinrich, J. Oster, C. Wang, Ö. Smedby, C. Bian, X. Yang, P.-A. Heng, A. Mortazi, U. Bagci, G. Yang, C. Sun, G. Galisot, J.-Y. Ramel, T. Brouard, Q. Tong, W. Si, X. Liao, G. Zeng, Z. Shi, G. Zheng, C. Wang, T. MacGillivray, D. Newby, K. Rhode, S. Ourselin, R. Mohiaddin, J. Keegan, D. Firmin, G. Yang, Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge, Medical Image Analysis 58 (2019) 101537. doi:10.1016/j.media.2019.101537.
[25] Z. Xiong, Q. Xia, Z. Hu, N. Huang, C. Bian, Y. Zheng, S. Vesal, N. Ravikumar, A. Maier, X. Yang, P.-A. Heng, D. Ni, C. Li, Q. Tong, W. Si, E. Puybareau, Y. Khoudli, T. Géraud, C. Chen, W. Bai, D. Rueckert, L. Xu, X. Zhuang, X. Luo, S. Jia, M. Sermesant, Y. Liu, K. Wang, D. Borra, A. Masci, C. Corsi, C. de Vente, M. Veta, R. Karim, C. J. Preetha, S. Engelhardt, M. Qiao, Y. Wang, Q. Tao, M. Nuñez-Garcia, O. Camara, N. Savioli, P. Lamata, J. Zhao, A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging, Medical Image Analysis 67 (2021) 101832. doi:10.1016/j.media.2020.101832.
[26] MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge, Synapse multi-organ segmentation dataset, in: MICCAI Challenge on Multi-Atlas Labeling, 2015, accessed: 2022-04-20.
URL https://www.synapse.org/#!Synapse:syn3193805/wiki/217789
[27] D. Gutman, N. C. F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC), https://arxiv.org/abs/1605.01397v1 (2016).
[28] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, A. Halpern, Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC), https://arxiv.org/abs/1710.05006v3 (2017).
[29] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, J. Rozeira, PH2 - A dermoscopic image database for research and benchmarking, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, pp. 5437–5440. doi:10.1109/EMBC.2013.6610779.
[30] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, E. Romero, An open access thyroid ultrasound image database, in: 10th International Symposium on Medical Information Processing and Analysis, Vol. 9287, SPIE, 2015, pp. 188–193. doi:10.1117/12.2073532.
[31] H. Gong, G. Chen, R. Wang, X. Xie, M. Mao, Y. Yu, F. Chen, G. Li, Multi-task learning for thyroid nodule segmentation with thyroid region prior, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 257–261. doi:10.1109/ISBI48211.2021.9434087.
[32] X. Huang, Z. Deng, D. Li, X. Yuan, MISSFormer: An Effective Medical Image Segmentation Transformer, https://arxiv.org/abs/2109.07162v2 (2021).
[33] Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, L. Li, Mamba-UNet: UNet-Like Pure Visual Mamba for Medical Image Segmentation (2024). arXiv:2402.05079, doi:10.48550/arXiv.2402.05079.
[34] Y. Ao, W. Shi, B. Ji, Y. Miao, W. He, Z. Jiang, MS-TCNet: An effective Transformer–CNN combined network using multi-scale feature learning for 3D medical image segmentation, Computers in Biology and Medicine 170 (2024) 108057. doi:10.1016/j.compbiomed.2024.108057.
[35] A. K. Titoriya, M. P. Singh, PVT-CASCADE network on skin cancer dataset, in: 8th International Conference on Computing in Engineering and Technology (ICCET 2023), Vol. 2023, 2023, pp. 480–486. doi:10.1049/icp.2023.1536.
[36] Collaborative networks of transformers and convolutional neural networks are powerful and versatile learners for accurate 3D medical image segmentation - ScienceDirect, https://www.sciencedirect.com/science/article/pii/S0010482523006935.
[37] W. Zhao, F. Wang, Y. Xie, Q. Wu, Y. Zhou, UD-Mamba: A pixel-level uncertainty-driven mamba model for medical image segmentation (2024).
[38] S. Sun, Q. She, Y. Ma, R. Li, Y. Zhang, LUCF-Net: Lightweight U-shaped Cascade Fusion Network for Medical Image Segmentation, https://arxiv.org/abs/2404.07473v1 (2024).
[39] M. M. Rahman, R. Marculescu, Multi-scale Hierarchical Vision Transformer with Cascaded Attention Decoding for Medical Image Segmentation (2023). arXiv:2303.16892.
[40] Y. Sadegheih, A. Bozorgpour, P. Kumari, R. Azad, D. Merhof, LHU-NET: A LIGHT HYBRID U-NET FOR COST-EFFICIENT, HIGH-PERFORMANCE VOLUMETRIC MEDICAL IMAGE SEGMENTATION.
[41] T. Y. Tsai, L. Lin, S. Hu, C. W. Tsao, X. Li, M.-C. Chang, H. Zhu, X. Wang, UU-Mamba: Uncertainty-aware U-Mamba for Cardiovascular Segmentation (2024). arXiv:2409.14305, doi:10.48550/arXiv.2409.14305.
[42] X. Lin, L. Yu, K.-T. Cheng, Z. Yan, BATFormer: Towards Boundary-Aware Lightweight Transformer for Efficient Medical Image Segmentation (2023). arXiv:2206.14409, doi:10.48550/arXiv.2206.14409.
[43] A. Tragakis, C. Kaul, R. Murray-Smith, D. Husmeier, The Fully Convolutional Transformer for Medical Image Segmentation, in: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 2023, pp. 3649–3658. doi:10.1109/WACV56688.2023.00365.
[44] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, PVT v2: Improved Baselines with Pyramid Vision Transformer, https://arxiv.org/abs/2106.13797v7 (2021). doi:10.1007/s41095-022-0274-8.
[45] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, https://arxiv.org/abs/1606.04797v1 (2016).
[46] Y. Cai, Y. Wang, MA-Unet: An improved version of Unet based on multi-scale and attention mechanism for medical image segmentation (2020). arXiv:2012.10952.
[47] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, Z. Wen, FAT-Net: Feature adaptive transformers for automated skin lesion segmentation, Medical Image Analysis 76 (2022) 102327. doi:10.1016/j.media.2021.102327.
[48] H. Pan, Q. Zhou, L. J. Latecki, SGUNET: Semantic Guided UNET For Thyroid Nodule Segmentation, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 630–634. doi:10.1109/ISBI48211.2021.9434051.
[49] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation (2016). arXiv:1511.00561, doi:10.48550/arXiv.1511.00561.
[50] J. Long, E. Shelhamer, T. Darrell, Fully Convolutional Networks for Semantic Segmentation, https://arxiv.org/abs/1411.4038v2 (2014).
[51] H. Gong, J. Chen, G. Chen, H. Li, F. Chen, G. Li, Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules, Computers in Biology and Medicine 106389 (2022) 1–12.
[52] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, X. Chen, CPFNet: Context Pyramid Fusion Network for Medical Image Segmentation, IEEE Transactions on Medical Imaging 39 (10) (2020) 3008–3018. doi:10.1109/TMI.2020.2983721.
[53] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation (2018). arXiv:1802.02611.
[54] S. Perera, P. Navard, A. Yilmaz, SegFormer3D: An Efficient Transformer for 3D Medical Image Segmentation (2024). arXiv:2404.10156, doi:10.48550/arXiv.2404.10156.
[55] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, nnFormer: In-

terleaved Transformer for Volumetric Segmentation (2022). arXiv:2109.03201, doi:10.48550/arXiv.2109.03201.

[56] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention U-Net: Learning Where to Look for the Pancreas, https://arxiv.org/abs/1804.03999v3 (2018).

[57] J. M. J. Valanarasu, V. M. Patel, UNeXt: MLP-based Rapid Medical Image Segmentation Network (2022). arXiv:2203.04967, doi:10.48550/arXiv.2203.04967.

[58] F. Tang, Q. Huang, J. Wang, X. Hou, J. Su, J. Liu, DuAT: Dual-Aggregation Transformer Network for Medical Image Segmentation (2022). arXiv:2212.11677, doi:10.48550/arXiv.2212.11677.

[59] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, D. Firmin, DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction, IEEE Transactions on Medical Imaging 37 (6) (2018) 1310–1321. doi:10.1109/TMI.2017.2785879.

[60] B. Yang, H. Peng, C. Guo, X. Luo, J. Wang, X. Long, SLP-Net:An efficient lightweight network for segmentation of skin lesions (2024). arXiv:2312.12789, doi:10.48550/arXiv.2312.12789.

[61] GFANet: Gated Fusion Attention Network for skin lesion segmentation, Computers in Biology and Medicine 155 (2023) 106462. doi:10.1016/j.compbiomed.2022.106462.

[62] D. Dai, C. Dong, S. Xu, Q. Yan, Z. Li, C. Zhang, N. Luo, Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation, Medical Image Analysis 75 (2022) 102293. doi:10.1016/j.media.2021.102293.

[63] S. Perera, Y. Erzurumlu, D. Gulati, A. Yilmaz, MobileUNETR: A Lightweight End-To-End Hybrid Vision Transformer For Efficient Medical Image Segmentation (2024). arXiv:2409.03062, doi:10.48550/arXiv.2409.03062.

[64] TC-Net: A joint learning framework based on CNN and vision transformer for multi-lesion medical images segmentation - ScienceDirect, https://www.sciencedirect.com/science/article/pii/S0010482523004328.

[65] J. Ruan, M. Xie, J. Gao, T. Liu, Y. Fu, EGE-UNet: An Efficient Group Enhanced UNet for skin lesion segmentation (2023). arXiv:2307.08473, doi:10.48550/arXiv.2307.08473.

[66] LeaNet: Lightweight U-shaped architecture for high-performance skin cancer image segmentation - ScienceDirect, https://www.sciencedirect.com/science/article/pii/S0010482524000039.

[67] VM-UNET-V2: Rethinking Vision Mamba UNet for Medical Image Segmentation, https://arxiv.org/html/2403.09157v1.

[68] W. Liao, Y. Zhu, X. Wang, C. Pan, Y. Wang, L. Ma, LightM-UNet: Mamba Assists in Lightweight UNet for Medical Image Segmentation (2024). arXiv:2403.05246, doi:10.48550/arXiv.2403.05246.

[69] C. Yuan, D. Zhao, S. S. Agaian, MUCM-Net: A Mamba Powered UCM-Net for Skin Lesion Segmentation (2024). arXiv:2405.15925, doi:10.48550/arXiv.2405.15925.

[70] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, D. Merhof, HiFormer: Hierarchical Multi-scale Representations Using Transformers for Medical Image Segmentation (2023). arXiv:2207.08518, doi:10.48550/arXiv.2207.08518.

[71] J. Ruan, S. Xiang, M. Xie, T. Liu, Y. Fu, MALUNet: A Multi-Attention and Light-weight UNet for Skin Lesion Segmentation (2022). arXiv:2211.01784, doi:10.48550/arXiv.2211.01784.

[72] E. K. Aghdam, R. Azad, M. Zarvani, D. Merhof, Attention Swin U-Net: Cross-Contextual Attention Mechanism for Skin Lesion Segmentation (2022). arXiv:2210.16898, doi:10.48550/arXiv.2210.16898.

[73] ULFAC-Net: Ultra-Lightweight Fully Asymmetric Convolutional Network for Skin Lesion Segmentation | IEEE Journals & Magazine | IEEE Xplore, https://ieeexplore.ieee.org/document/10077446.

[74] S. Tang, C. F. Cheang, X. Yu, Y. Liang, Q. Feng, Z. Chen, TransCS-Net: A hybrid transformer-based privacy-protecting network using compressed sensing for medical image segmentation, Biomedical Signal Processing and Control 86 (2023) 105131. doi:10.1016/j.bspc.2023.105131.

Declaration of interests

☐The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Fei Yang reports financial support was provided by the Natural Science Foun- dation of Shandong Province. Fei Yang reports financial support was provided by the National Natural Science Foundation of China. Fei Yang reports financial support was provided by Guangdong Basic and Applied Basic Research Foundation. Fei Yang reports equipment, drugs, or supplies was provided by NVIDIA Corp. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.