CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

A Lightweight Hybrid Network for Medical Image Segmentation With Adaptive Feature Selection

Zhouwei Lin[1] | Yudong Xia[2] | Pingping Wu[3] | Han Lin[3] | Qingsong Xu[4] | Zheng Fan[5] | Juan Tu[6] | Jun Wu[7]

[1]School of Mechanics and Engineering Science, Peking University, Beijing, China | [2]Department of Cardiology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China | [3]School of Engineering Audit, Nanjing Audit University, Nanjing, China | [4]Department of Electromechanical Engineering, Faculty of Science and Technology, University of Macau, Taipa, Macau, China | [5]School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore | [6]Key Laboratory of Modern Acoustics (MOE), Department of Physics, Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing, China | [7]Department of Geriatric Cardiology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China

Correspondence: Jun Wu (wujun9989@njmu.edu.cn)

Received: 8 April 2026 | **Revised:** 18 June 2026 | **Accepted:** 1 July 2026

Keywords: biomedical engineering | deep learning | machine learning | medical imaging | segmentation | vision recognition

ABSTRACT

Accurate medical image segmentation with low model complexity remains difficult because lesions are often small in scale and boundary cues are easily corrupted by noise. Although recent segmentation methods have achieved strong performance, many of them rely on increasingly complex architectures with high computational costs, limiting their applicability in resource-constrained settings. To alleviate this issue, we propose PLNet, a lightweight hybrid segmentation framework for medical image analysis. The network improves representation learning by integrating fine-grained local-structure perception with long-range contextual modelling, thereby enhancing the representation of complex medical images. In addition, a feature selection mechanism is employed to emphasise informative feature responses and suppress redundant activations while preserving discriminative representations. Extensive experiments on multiple datasets demonstrate that PLNet achieves competitive IoU and Dice scores with relatively low model complexity. Overall, these results highlight the potential of PLNet as an effective lightweight framework that balances segmentation performance and computational complexity.

1 | Introduction

Medical image segmentation is a key component of computer-aided diagnosis and clinical decision support, enabling precise localization of anatomical structures and pathological regions for effective disease analysis, treatment planning, and outcome evaluation. Recent advances in deep learning-based models have led to strong performance across a wide range of medical imaging benchmarks. Nevertheless, accurately segmenting lesions with small lesion-to-image area ratios, low contrast, and substantial imaging noise remains challenging. These difficulties mainly arise from background dominance, blurred lesion boundaries, and the inherent noise and variability in medical images, which together increase segmentation uncertainty [1–4].

From a modelling perspective, these challenges are further intensified under limited or sparsely annotated training data. Data-driven approaches depend heavily on sufficient supervision to learn discriminative representations, and their performance often degrades significantly in weakly supervised scenarios. Classical convolutional neural network (CNN)-based architectures, such as U-Net [5], nnU-Net [6], and V-Net [7], often introduce multi-scale context extraction modules, such as CE-Net

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

[8], to enhance the representation of local features. Nevertheless, constrained by their limited receptive fields, CNN-based models have difficulty in modelling global dependencies and long-range contextual relationships.

To address this limitation, the Vision Transformer (ViT) [9] was introduced, demonstrating strong capability in modelling global context and long-range dependencies. Building on this paradigm, Transformer-based segmentation models such as Swin-UNet [10] and CSWin Transformer [11] have been developed. Swin-UNet employs non-overlapping local windows with a shifted-window strategy to balance global modelling and computational efficiency, while CSWin Transformer further improves global dependency modelling by adopting cross-shaped attention along horizontal and vertical directions. By explicitly modelling contextual relationships through self-attention, these Transformer-based architectures effectively compensate for the limited long-range modelling capability of CNNs, leading to enhanced global semantic consistency in medical image segmentation.

Although the ViT encoder is effective at capturing global semantic context, its self-attention mechanism may overlook fine-grained local details. To mitigate this limitation, a number of hybrid architectures have been proposed that combine CNNs with Transformers, such as TransUNet [12] and LeViT-UNet [13]. By combining CNNs and Transformers, existing approaches seek to jointly model local details and long-range contextual information for medical image segmentation. However, effective fusion of local and global features remains difficult. Preserving feature consistency across multiple encoders while fully exploiting their complementary characteristics is challenging, and such hybrid architectures often incur increased computational and memory costs, limiting scalability and practical deployment. In addition, Transformer-based mechanisms alone may still fail to capture sufficiently rich long-range context in complex scenarios, leading to incomplete or inaccurate boundary delineation.

Structured State Space Sequence Models (SSM) [14] have recently emerged as an efficient paradigm for long-sequence modelling, demonstrating strong capability in capturing long-range dependencies. Unlike ViT architectures, SSMs achieve linear or near-linear complexity with increasing sequence length, while retaining effective long-term dependency modelling, resulting in strong performance on long-sequence analysis tasks. In the visual domain, Liu proposed VMamba [15], which incorporates the Mamba module and a cross-scan mechanism to alleviate direction sensitivity caused by converting 2D images into 1D sequences. In medical image segmentation, models such as U-Mamba [16], SegMamba [17], and Swin-UMamba [18] adapt state-space modelling to segmentation through different CNN–Mamba, encoder–decoder, or Swin–Mamba designs. Although these models reduce the computational cost of modelling long-range dependencies, the sequence scanning strategy used in Mamba-based networks limits their ability to perceive local features, resulting in weaker fine-detail representation compared with the local receptive fields of CNN and the small-window attention mechanism of Vision Transformers. In addition, most Mamba-based segmentation networks are trained from scratch without effective pretraining, which is particularly challenging for medical imaging tasks with limited data, as it increases the likelihood of overfitting. From a practical perspective, Mamba-based models still lack extensive empirical validation, and their effectiveness, robustness, and computational efficiency in real-world applications remain uncertain, restricting their broader adoption in medical image analysis. Recently, large-model–based approaches, including DINOv3 [19], MedSAM [20], and SAM-Adapter [21, 22], have demonstrated encouraging performance in medical image segmentation. Despite their effectiveness, several limitations remain. MedSAM requires user-provided bounding-box prompts during inference, thereby introducing additional manual interaction, whereas SAM-Adapter requires task-specific adapter training, which increases the training cost and may limit its deployment in resource-constrained environments.

In this paper, we propose PLNet, a lightweight hybrid framework that balances segmentation accuracy, model size, and computational complexity for medical image segmentation. The core innovation of PLNet lies in its overall collaborative design, which integrates multi-branch feature extraction, adaptive feature selection, and lightweight decoding into a unified segmentation framework. Motivated by the effectiveness of global-local representation learning, PLNet constructs a complementary encoder architecture consisting of three branches: a Convolutional Neural Network (CNN) branch for local detail extraction, a lightweight Vision Transformer (ViT) branch for global semantic context modelling, and a Visual State Space Model (VSSM) branch for structured spatial dependency modelling, as shown in Figure 1. Rather than simply stacking existing CNN, ViT, and Mamba modules, PLNet assigns different roles to the three branches and introduces targeted lightweight designs to enhance their cooperation. The CNN branch uses an MSCA module to strengthen local texture and boundary representation, the ViT branch combines max pooling and self-attention to reduce computational cost while preserving global semantic modelling, and the VSSM branch introduces diagonal scanning directions to improve structured spatial dependency modelling. Furthermore, a sparse Mixture-of-Experts (MoE) fusion module is introduced to integrate heterogeneous encoder features by selectively activating and aggregating expert outputs, thereby emphasising informative representations and suppressing redundant feature responses. Through this coordinated design, PLNet achieves complementary feature extraction and low-complexity feature aggregation, leading to a favourable balance between segmentation accuracy, model size, and computational complexity. Overall, the main contributions of this work are summarised as follows:

- We propose PLNet, a lightweight hybrid segmentation framework for medical image segmentation. Its core design is a coordinated three-branch encoder, referred to as the CVT encoder in this work, where CVT denotes the collaborative integration of CNN, ViT, and VSSM branches. Rather than simply stacking existing modules, the CNN, ViT, and VSSM branches are assigned complementary roles to capture local texture and boundary details, global semantic context, and structured spatial dependencies, respectively.
- We introduce targeted lightweight designs for different encoder branches and adopt a sparse Mixture-of-Experts (MoE)-based fusion module to integrate heterogeneous features from different encoders adaptively. The MoE module emphasises informative representations and suppresses

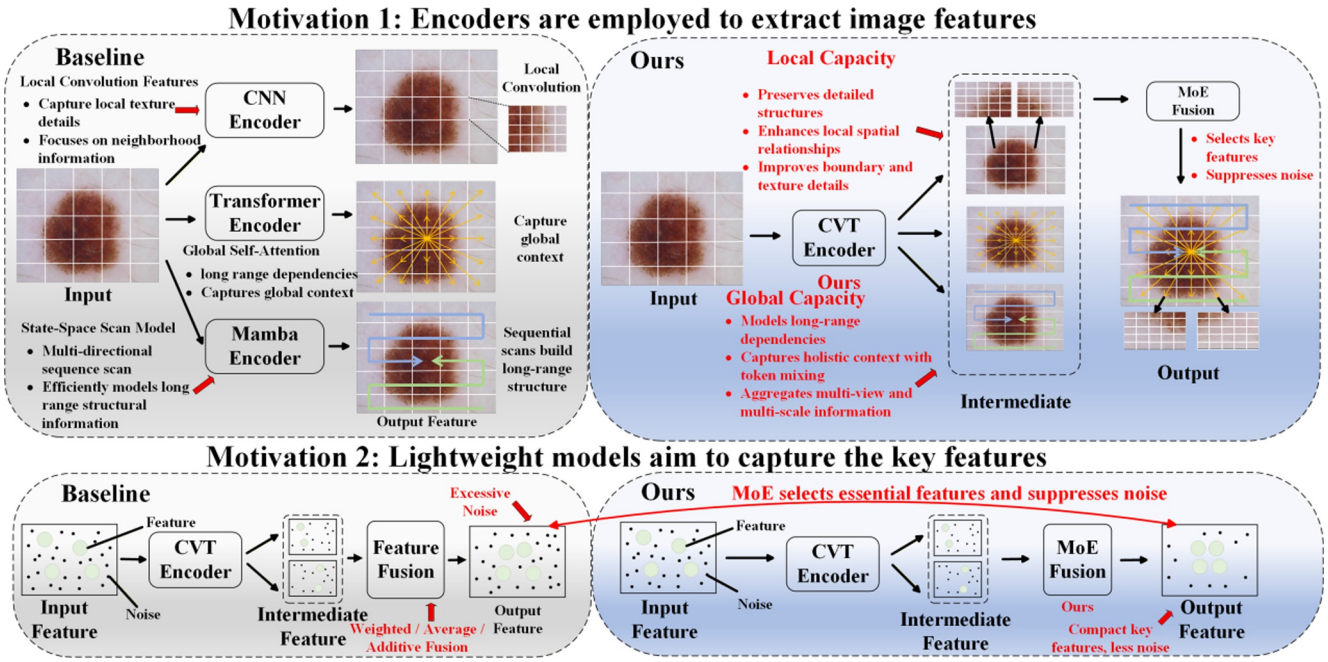


FIGURE 1 | Overview of the proposed CVT encoder framework and lightweight feature selection mechanism. Here, CVT denotes the CNN-ViT-VSSM triple-branch encoder in PLNet. Motivation 1 shows the complementary roles of different encoders: the CNN branch captures local texture and boundary details, the ViT branch models global semantic context through self-attention, and the VSSM branch captures structured spatial dependencies through multi-directional scanning. The CVT encoder integrates these complementary representations through token mixing and multi-scale aggregation, while the MoE fusion module selects informative features and suppresses redundant responses. Motivation 2 illustrates that the proposed feature selection mechanism reduces noisy and redundant responses, leading to more compact and discriminative representations.

redundant responses, while the lightweight MLP decoder serves as a low-complexity prediction head.

- Extensive experiments on multiple medical image datasets demonstrate that PLNet achieves competitive segmentation performance with relatively low FLOPs and model size, confirming its effectiveness and lightweight design for medical image segmentation.

2 | Related Work

2.1 | Convolutional Neural Network-Based Segmentation

The emergence of convolutional neural networks (CNNs) has led to substantial advances in computer vision. Their powerful feature extraction capability based on convolutional operations has led to significant advances in image classification, object detection, and semantic segmentation. CNN-based segmentation techniques have also been rapidly extended to the field of medical imaging, playing a crucial role in tumour detection and organ segmentation. The U-Net model proposed by Ronneberger et al. [5] laid the foundation for medical image segmentation. Based on its U-shaped architecture, researchers have introduced various improvements. Several U-Net variants enhance feature fusion through improved skip-connection designs. UNet++ [1] introduces dense connections inspired by DenseNet [23] to strengthen feature representation and mitigate detail loss during downsampling, while UNet-3+ [24] further integrates multi-scale features via full-scale skip connections and deep supervision. In addition, U-Net v2 [25], Rethinking U-Net [26], and MPR-Net [27] refine feature representation by injecting semantic information and enhancing details through attention-based mechanisms. The latter proposes Task-Adaptive Mixture of Skip Connections (TA-SC), which employs learnable dynamic weights to adaptively fuse features from different layers, allowing the network to balance low-level detail and high-level semantics according to the task. With the emergence of the Kolmogorov–Arnold Network (KAN), U-KAN [28] and ResU-KAN [29] models have been developed on the U-Net framework. Their core idea is to replace traditional convolutional or MLP layers with KAN layers, enabling interpretable nonlinear feature modelling within the U-Net architecture and constructing an efficient, lightweight, and theoretically interpretable backbone for medical image segmentation. Although these models can effectively extract local features, their reliance on limited receptive fields makes it difficult to capture global contextual information. This limitation becomes particularly evident when dealing with medical images that involve complex structures or long-range dependencies, thereby constraining the further improvement of segmentation accuracy.

2.2 | Vision Transformers for Medical Image Analysis

With the success of Transformers in natural language processing, their capability to model long-range dependencies has drawn increasing interest in computer vision. This led to the development of the Vision Transformer (ViT) [9], which models images as sequences of patch embeddings to capture global context. However, the high data demand and computational cost of ViT limit its practical deployment. To mitigate these issues, DeiT [30] introduces a convolution-based teacher–student distillation strategy to improve training efficiency.

Building upon ViT, a variety of hybrid CNN–Transformer architectures have been proposed to jointly leverage local feature extraction and global context modelling. Representative examples include Swin-Unet [10], which adopts a hierarchical Swin Transformer for efficient global representation, and TransUNet [12], which combines a CNN backbone with Transformer modules. Subsequent models such as GH-UNet [31], BEFUnet [32], and uncertainty-driven reliable dynamic fusion [33] further enhance multi-level feature fusion through specialised fusion mechanisms.

Although these hybrid designs improve segmentation accuracy by jointly modelling local and global representations, they remain computationally expensive. This is mainly due to the quadratic complexity of global self-attention in ViT and the increased parameter size introduced by dual CNN–Transformer encoders. Such limitations are particularly pronounced in dense medical image segmentation tasks, posing challenges for efficiency and practical deployment.

2.3 | Long-Range Context Modelling in Medical Images via Visual State Spaces

To address the challenge of modelling long-range dependencies with reduced computational cost, Gu proposed the State Space Model (S4) [14], which enables efficient long-sequence modelling while preserving global semantic information. This line of work was further advanced with the introduction of Mamba (S6) [34], which incorporates an input-dependent selection mechanism and hardware-aware design to improve efficiency and reduce parameter overhead.

In the vision domain, Liu adapted Mamba to visual tasks by proposing VMamba [15], introducing a Cross-Scan Module to transform image data into ordered sequences. Building on this foundation, several Mamba-based segmentation models have been developed. For example, SegMamba [17] adopts a CNN–SSM hybrid architecture to replace Transformer-based attention for long-range dependency modelling, while Swin-UMamba [18] integrates SSM into the Swin-Transformer framework, achieving efficient long-sequence processing with reduced computational cost. These approaches demonstrate the effectiveness of state-space models in improving both accuracy and efficiency for medical image segmentation.

2.4 | Parallel Network Frameworks for Semantic Segmentation

In 2019, Chen et al. proposed Collaborative Global-Local Networks (GLNet) [35], which contains a global branch for capturing contextual information from downsampled images and a local branch for extracting fine-grained details from cropped patches. The two branches are integrated through deep feature sharing and feature map regularisation to produce high-quality segmentation results. Based on this global-local paradigm, CascadePSP [36] adopts a coarse-to-fine refinement strategy to recover global structures and refine local details, while LEFormer [37] combines CNN-based local feature extraction with Transformer-based long-range semantic modelling. Recent medical image segmentation studies have further extended global-local representation learning to lesion segmentation and multi-level feature fusion. GLMKD [38] introduces joint global and local mutual knowledge distillation for weakly supervised lesion segmentation in histopathology images, improving lesion localization by combining global contextual cues and multi-scale local details. GLAD-Net [39] proposes global-local adaptive fusion and cross-stage distillation for cross-level multi-scale medical image segmentation, enhancing hierarchical feature interaction and reducing semantic inconsistency. These studies demonstrate the effectiveness of integrating global context, local details, and multi-level features. In contrast, PLNet constructs a lightweight hybrid framework by combining CNN, ViT, and VSSM branches with sparse MoE-based adaptive feature fusion.

3 | Methodology

3.1 | Overview

In this paper, we propose PLNet, a lightweight semantic segmentation model for medical image segmentation. As shown in Figure 2, the overall pipeline follows three stages: multi-branch feature encoding, feature fusion among different encoders, and lightweight decoding. In the encoding stage, PLNet employs three parallel branches, namely CNN, ViT, and VSSM, as detailed in Section 3.2. The CNN branch captures local texture and boundary details, the ViT branch models global semantic context through token interaction, and the VSSM branch captures structured spatial dependencies through multi-directional scanning, thereby providing complementary multi-scale representations. In the fusion stage, the MoE module adaptively integrates features from different encoder branches by selecting suitable experts to emphasise informative representations and reduce redundant responses, as described in Section 3.3. In the decoding stage, a lightweight decoder projects multi-stage features into a unified channel space using MLP layers and restores the final segmentation map to the original resolution through upsampling, as described in Section 3.4.

3.2 | Encoder

3.2.1 | Convolutional Neural Network Encoder

PLNet consists primarily of three parallel encoder branches. The CNN encoder is primarily composed of multiple CNN layers, where each CNN layer includes two components: a DW Layer [40] and an MSCA Layer [37]. The input data passes through the DW Layer, where different convolution kernels, padding, and strides are applied to generate multi-level feature maps. In PLNet, the encoder generates hierarchical feature maps at resolutions of $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$.

Additionally, we employ the MSCA Layer. As shown in Figure 3, The MSCA module [37] combines dilated convolution [41] and CBAM [42], leveraging multi-scale feature extraction and attention mechanisms to enhance feature representation. First, dilated convolutions with the same kernel size but different dilation rates are employed to extract feature maps at multiple scales. Feature maps with smaller dilation rates emphasise edge and fine-grained details, whereas those with larger dilation rates capture more global contextual information. In the channel attention module, max pooling and average pooling are applied to each input feature map to obtain the max-pooled feature $F^{C}_{\text{Max}} \in \mathbb{R}^{C\times 1\times 1}$ and the

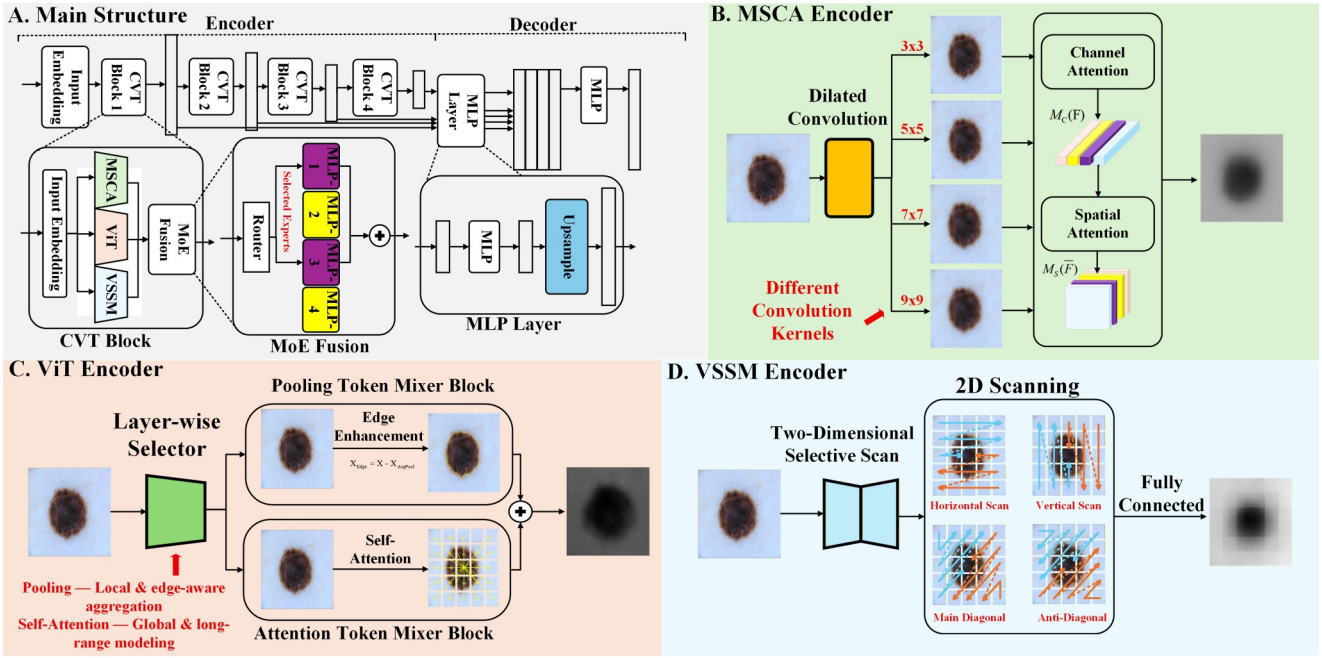


FIGURE 2 | Detailed architecture of PLNet. (A) Main structure: PLNet adopts a pyramid CVT encoder to extract multi-level features, uses a sparse MoE module for adaptive feature fusion, and employs a lightweight decoder to generate the final segmentation mask. (B) MSCA encoder: Channel attention and spatial attention are incorporated to enhance local texture and boundary feature extraction. (C) ViT encoder: A selective token mixing design is introduced to reduce computational complexity while preserving global semantic modelling capability. (D) VSSM encoder: Based on SS2D, the VSSM encoder adopts eight directional scanning operations to enhance structured spatial dependency modelling.

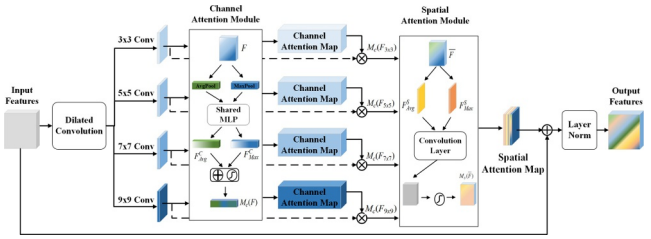


FIGURE 3 | The MSCA module captures multi-level feature representations by employing convolution kernels with varying receptive fields and further strengthens these features using channel and spatial attention mechanisms.

average-pooled feature $F^C_{\text{Avg}} \in \mathbb{R}^{C\times1\times1}$. These two feature representations are subsequently fed into a shared network consisting of a multilayer perceptron (MLP) with hidden layers, producing the channel attention map $M_c \in \mathbb{R}^{C\times1\times1}$. The specific equations are given as follows:

$$\begin{aligned}\mathbf{M}_c(\mathbf{F}) &= \sigma(\text{MLP}(\text{AvgPool}(\mathbf{F})) + \text{MLP}(\text{MaxPool}(\mathbf{F}))),\\ \mathbf{F}_c &= \mathbf{M}_c(\mathbf{F}) \odot \mathbf{F},\end{aligned} \tag{1}$$

where σ denotes the sigmoid function, $\odot$ represents element-wise multiplication, and $\mathbf{F}_c$ denotes the channel-refined feature map. In the spatial attention module, average pooling and max pooling are applied along the channel dimension of $\mathbf{F}_c$ to obtain $\mathbf{F}^s_{\text{Avg}} \in \mathbb{R}^{1\times H\times W}$ and $\mathbf{F}^s_{\text{Max}} \in \mathbb{R}^{1\times H\times W}$, respectively. These feature maps are concatenated along the channel dimension and processed by a convolutional layer to generate the spatial attention map $\mathbf{M}_s \in \mathbb{R}^{1\times H\times W}$. The corresponding operation is defined as follows:

$$\begin{aligned}\mathbf{M}_s(\mathbf{F}_c) &= \sigma\left(f^{(3\times3)}\left(\left[\text{AvgPool}_c(\mathbf{F}_c); \text{MaxPool}_c(\mathbf{F}_c)\right]\right)\right),\\ \mathbf{F}_s &= \mathbf{M}_s(\mathbf{F}_c) \odot \mathbf{F}_c,\end{aligned} \tag{2}$$

where $f^{(3\times3)}$ denotes a convolutional operation with a kernel size of 3×3, the subscript c indicates pooling along the channel dimension, and $\mathbf{F}_s$ denotes the spatially refined feature map. By sequentially applying channel and spatial attention, the MSCA module emphasises informative channel responses and spatial regions, thereby strengthening local texture and boundary representations while suppressing redundant feature activations.

3.2.2 | *Vision Transformer Encoder*

Since the CNN encoder mainly focuses on local feature extraction, it is limited in modelling long-range semantic dependencies. To address this issue, a Vision Transformer (ViT) encoder is introduced to enhance global feature representation [9]. To maintain consistency with the pyramid representation of the CNN branch, the ViT encoder adopts a Pyramid Vision Transformer (PVT) structure, generating multi-scale feature maps with the same spatial resolutions as those of the CNN encoder [43]. As illustrated in Figure 4, adjacent patches are progressively merged to form a pyramid-structured Transformer.

However, the self-attention operation in ViT has a computational complexity of $O(N^2)$, which becomes expensive for high-resolution medical images. The attention mechanism is defined as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_{\text{head}}}}\right)V, \tag{3}$$

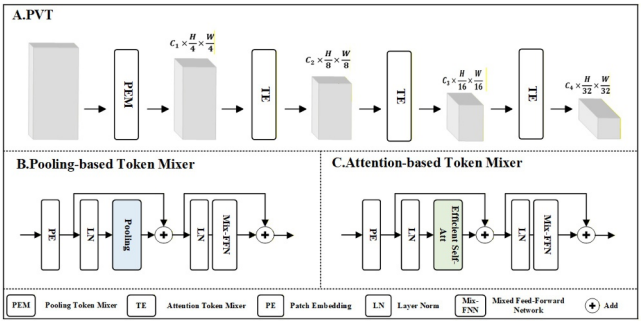


FIGURE 4 | The main structures of the ViT encoder are as follows: (A) PVT: A Pyramid Vision Transformer structure. (B) Pooling-based Token Mixer: Replaces the self-attention operation with a lightweight pooling mechanism to aggregate local spatial information with lower token-interaction cost. (C) Attention-based Token Mixer: Uses self-attention to capture long-range dependencies across tokens, enhancing the model's ability to integrate global information. All modules include patch embedding (PE), layer normalisation (LN), residual connections, and a mixed feed-forward network (Mix-FFN).

where N denotes the sequence length, C denotes the feature dimension, and d_{head} is the dimension of each attention head. To reduce the computational burden, a reduction rate R is introduced to shorten the sequence length, as follows:

$$\begin{aligned} K_r &= \text{Reshape}\left(\frac{N}{R}, C{\cdot}R\right)(K), \\ \widehat{K} &= \text{Linear}\left(C_{\text{input}}{\cdot}R,\ C_{\text{output}}\right)(K_r). \end{aligned} \tag{4}$$

Here, $K \in \mathbb{R}^{N\times C}$ denotes the original input sequence. It is first reshaped into $K_r \in \mathbb{R}^{\frac{N}{R}\times(C\cdot R)}$, which shortens the sequence length while aggregating channel information, and is then linearly projected to obtain the reduced sequence $\widehat{K}$. This operation reduces the computational complexity from $O(N^2)$ to $O\left(\frac{N^2}{R}\right)$. To align the ViT branch with the multi-scale feature maps generated by the CNN encoder, the reduction rates are set to (8, 4, 2, 1).

In addition, conventional positional encodings are resolution-dependent and usually require interpolation when the input resolution changes during inference, which may impair model generalisation. Therefore, Mix-FFN [44] is adopted to implicitly encode positional information by integrating a 3 × 3 depthwise convolution (DWConv) and an MLP into the FFN, thereby eliminating the dependence on explicit positional encodings. Unlike a standard convolution, the DWConv operates independently on each channel, allowing local spatial information to be introduced with only a small number of additional parameters. This design is consistent with the lightweight motivation of PLNet. The formulation is given as follows:

$$x_{\text{out}} = \text{MLP}_2(\text{GELU}(\text{DWConv}_{3\times3}(\text{MLP}_1(x_{\text{in}})))) + x_{\text{in}}. \tag{5}$$

Although the PVT-style reduction effectively shortens the token sequence, the ViT branch still introduces non-negligible computational overhead. To reduce the token-interaction cost, we redesign the ViT encoder with a lightweight strategy inspired by MetaFormer, PoolFormer [45], and LEFormer [37]. As shown in Figure 4c, the Transformer is retained as the basic framework, while different Token Mixer modules are introduced for lightweight token interaction. In addition, the original MLP block is replaced with Mix-FFN to better integrate spatial and channel information. Specifically, a Pooling-based Token Mixer is employed in the first stage to aggregate local context with lower token-interaction cost, whereas Attention-based Token Mixers are used in the remaining three stages to preserve long-range dependency modelling. In summary, the PVT-style reduction is used for sequence compression, while the Pooling/Attention token mixers are used for lightweight token interaction.

3.2.3 | *Visual State Space Model Encoder*

To complement the local modelling ability of the CNN encoder and the long-range dependency modelling ability of the ViT encoder, we introduce a lightweight VSSM encoder to further capture spatially ordered contextual representations. Specifically, the CNN encoder focuses on local texture and boundary cues, while the ViT encoder models global semantic dependencies through self-attention. Considering that these two encoders already account for a large proportion of the computational cost, the additional encoder should capture complementary contextual information without substantially increasing model complexity. Motivated by State Space Models, we incorporate the VSSM encoder [15] into the proposed architecture. VSSM provides an efficient way to model long-range spatial dependencies by projecting input sequences into hidden states and propagating contextual information along spatial directions. The corresponding formulation is given as follows:

$$\begin{aligned} h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \\ y_t &= \mathbf{D}h_t + \mathbf{F}x_t, \end{aligned} \tag{6}$$

where x_t represents the one-dimensional input sequence at time step t, and h_t denotes the corresponding hidden state. $\overline{\mathbf{A}}$ is the discretised state transition matrix that governs the dynamic evolution of the hidden state, while $\overline{\mathbf{B}}$ is the discretised input matrix that maps the input x_t into the state space. In the output equation, y_t denotes the output at time step t, $\mathbf{D}$ is the output matrix that projects the hidden state h_t to the output space, and $\mathbf{F}$ represents the direct input-to-output mapping. Since State Space Models are originally formulated as continuous-time systems, discretisation is required to better integrate them into deep learning frameworks. In the Mamba model, the continuous matrices $\mathbf{A}$ and $\mathbf{B}$ are discretised into $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ using zero-order hold (ZOH) discretisation with a step-size parameter Δ, as expressed below:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}[\exp(\Delta\mathbf{A}) - \mathbf{I}]\Delta\mathbf{B}. \end{aligned} \tag{7}$$

Assuming a zero initial state, the discretised SSM can be computed either through linear recurrence or an equivalent global convolution. The convolutional form is given by:

$$\begin{aligned} y &= x * \overline{\mathbf{K}} + \mathbf{F}x, \\ \overline{\mathbf{K}} &= \left(\mathbf{D}\overline{\mathbf{B}}, \mathbf{D}\overline{\mathbf{A}}\overline{\mathbf{B}}, \ldots, \mathbf{D}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}\right), \end{aligned} \tag{8}$$

where $\overline{\mathbf{K}}$ denotes the structured convolution kernel, L is the sequence length, and $*$ denotes the convolution operation.

In the VSSM encoder, the input feature map $x \in \mathbb{R}^{C\times H\times W}$ is partitioned into non-overlapping 4×4 patches and linearly projected to $\bar{x} \in \mathbb{R}^{C\times H/4\times W/4}$. After layer normalisation, the features are divided into two branches: a residual branch for feature fusion and a core VSSM branch. In the core branch, a linear layer, depthwise convolution, and SiLU activation are first applied to capture local spatial information, followed by the SS2D module to model long-range dependencies and structured spatial relationships.

The SS2D module consists of scan expansion, an S6 block, and scan merging. As illustrated in Figure 5, scan expansion converts the feature map into multi-directional sequences along predefined axial and diagonal paths. These sequences are selectively processed by the S6 module [34], and subsequently merged through additive fusion to reconstruct the spatial representation. After normalisation, the two branches are combined via element-wise addition, followed by a linear projection to produce the output feature map.

Additionally, the PLNet encoder is composed of four CVT layers. In the VSSM branch, the Patch Merging strategy [46] is employed to reduce the spatial resolution of the input feature maps while increasing the channel dimension. Since the feature maps generated by the other two encoder branches have different spatial sizes, Patch Merging is also used to produce feature maps with aligned resolutions, thereby facilitating feature fusion. In the VSSM encoder, (2, 2, 4, 2) VSSM blocks are adopted across the four stages, with the number of channels at each stage set to (C, $2C$, $4C$, $6C$), where the default value of C is 32. More implementation details can be found in Table 1.

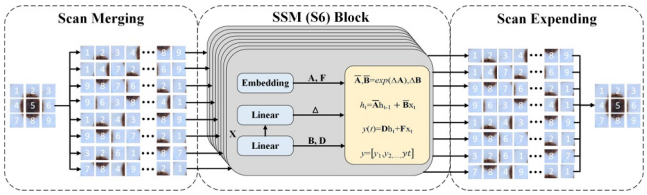


FIGURE 5 | The structure of SS2D: The input feature map is first partitioned into patches, which are subsequently flattened along eight distinct scanning paths (Cross-Scan). These sequences of image patches are then individually processed by separate S6 blocks. Following this, the outputs are merged to reconstruct the 2D feature map, which serves as the final output (Cross-Merge).

3.3 | Mixture-of-Experts Model

Since the proposed model employs three complementary encoders, the fusion stage needs to integrate heterogeneous features from different representation spaces. Specifically, the CNN encoder mainly captures local texture and boundary cues, the ViT encoder models long-range semantic dependencies, and the VSSM encoder provides spatially ordered contextual representations. At each encoder stage, the feature maps generated by the CNN, ViT, and VSSM branches are first aligned to the same spatial resolution and concatenated along the channel dimension. The concatenated feature is then fed into a Mixture-of-Experts (MoE) module for adaptive feature processing and fusion. By selectively processing the jointly represented multi-source features, the MoE module enables the model to emphasise discriminative responses while suppressing redundant or less informative feature activations.

Specifically, we adopt a sparse MoE structure in which a gating network predicts a weight distribution over the experts, and only the top-K experts are dynamically activated for each input. The outputs of the selected experts are then combined through a weighted sum. Unlike dense expert fusion, this strategy avoids activating all experts simultaneously, thereby reducing per-input computation and redundant feature processing. The gating mechanism is formulated as follows:

$$g = G(x) = \mathrm{Softmax}\left(W_g \cdot x + \epsilon\right), \tag{9}$$

where x denotes the channel-wise concatenated feature from the CNN, ViT, and VSSM branches, W_g denotes the weight matrix of the gating network, and ϵ represents the noise term used in Noisy Top-K Gating to prevent the gating network from prematurely converging to only a few experts. The Softmax operation ensures that the gating weights sum to 1, producing a dense gating vector $g = [g_1, g_2, ..., g_n]$, where n denotes the number of experts.

To obtain sparse expert activation, the elements in g are sorted, and only the Top-K gating weights are retained. Let $\mathcal{S}_K$ denote the index set of the selected Top-K experts. The sparse gating coefficient $\tilde{g}_i$ is defined as:

$$\tilde{g}_i = \begin{cases} g_i, & i \in \mathcal{S}_K, \\ 0, & i \notin \mathcal{S}_K. \end{cases} \tag{10}$$

The selected sparse gating weights are then normalised as:

TABLE 1 | The parameter settings for each encoder in the PLNet model include patch size (K_i), stride (S_i), padding size (P_i), output channel (C_i), reduction ratio (R_i), head number (N_i), number of encoder blocks (L_i) and depths (D_i).

Stage	Output size	Output channel	CNN encoder	ViT encoder	VSSM encoder
1	$\frac{H}{4} \times \frac{W}{4}$	96	$K_1 = 7$; $S_1 = 4$; $P_1 = 3$; $C_1 = 32$	$R_1 = 8$; $N_1 = 1$; $L_1 = 2$; $C_1 = 32$	$K_1 = 4$; $D_1 = 2$; $C_1 = 32$
2	$\frac{H}{8} \times \frac{W}{8}$	192	$K_2 = 3$; $S_2 = 2$; $P_2 = 1$; $C_2 = 64$	$R_2 = 4$; $N_2 = 2$; $L_2 = 2$; $C_2 = 64$	$K_2 = 4$; $D_2 = 2$; $C_2 = 64$
3	$\frac{H}{16} \times \frac{W}{16}$	384	$K_3 = 3$; $S_3 = 2$; $P_3 = 1$; $C_3 = 128$	$R_3 = 2$; $N_3 = 5$; $L_3 = 2$; $C_3 = 128$	$K_3 = 4$; $D_3 = 4$; $C_3 = 128$
4	$\frac{H}{32} \times \frac{W}{32}$	576	$K_4 = 3$; $S_4 = 2$; $P_4 = 1$; $C_4 = 192$	$R_4 = 1$; $N_4 = 6$; $L_4 = 3$; $C_4 = 192$	$K_4 = 4$; $D_4 = 2$; $C_4 = 192$

$$w_i = \frac{\tilde{g}_i}{\sum_{j \in \mathcal{S}_K} \tilde{g}_j}, \quad i \in \mathcal{S}_K, \tag{11}$$

where w_i denotes the normalised weight of the i-th selected expert. The final output is obtained by aggregating only the selected Top-K experts:

$$y = \sum_{i \in \mathcal{S}_K} w_i \cdot E_i(x), \tag{12}$$

where E_i denotes the i-th expert network. Although all expert parameters remain part of the model, only the selected subset of experts participates in the forward computation for each input. Therefore, sparse MoE does not necessarily reduce the total number of model parameters; instead, it reduces per-input computation and avoids redundant feature processing. This mechanism enables the model to adaptively aggregate complementary encoder features while emphasising informative boundary cues, texture patterns, and high-level semantic representations.

3.4 | Lightweight Decoder

Since PLNet is designed to maintain a compact model size and low computational complexity, we adopt a lightweight decoder based on an MLP prediction head. Inspired by SegFormer [44], the decoder is constructed with an MLP-based prediction head instead of a heavy convolutional decoder. Although the decoder is lightweight, the final prediction is not generated from a single high-level feature map. Instead, multi-stage features from the hierarchical encoder are first aggregated and aligned before prediction. In this process, shallow features provide fine-grained spatial and boundary cues, while deeper features contribute semantic context and a larger effective receptive field. Therefore, the decoder mainly serves as a lightweight projection and aggregation head, while boundary-aware and semantic representations are enhanced before decoding through the multi-branch encoder and MoE-based feature selection module.

As depicted in Figure 2, the features from the CNN, ViT, and VSSM branches are concatenated along the channel dimension at each stage, producing feature maps with the size of $(C_{\text{CNN}} + C_{\text{ViT}} + C_{\text{VSSM}}) \times \frac{H}{2^{(i+1)}} \times \frac{W}{2^{(i+1)}}, \quad i \in \{1, 2, 3, 4\}$. The concatenated feature at each stage is subsequently processed by the sparse MoE module to obtain an adaptively fused representation. The resulting multi-stage features are then projected into a unified channel space through MLP layers and resized to one-quarter of the input resolution. These resized features are further aggregated to form a global representation, which is used to predict the segmentation map $\mathbf{M} \in \mathbb{R}^{N_{\text{cls}} \times \frac{H}{4} \times \frac{W}{4}}$. Finally, the prediction is restored to the original image resolution through bilinear upsampling.

4 | Experiments

4.1 | Datasets

To systematically evaluate the effectiveness and robustness of PLNet, experiments were conducted on multiple medical image segmentation datasets with varied characteristics and levels of difficulty. Given that lesion regions in medical images often occupy a limited area of the image, two representative skin cancer datasets, ISIC2017 [47] and ISIC2018 [48, 49] ($train : test = 9 : 1$), were selected to examine model performance under small-target segmentation scenarios. To further examine the ability of the model to handle noise, we included the BUSI [50] ultrasound dataset ($train : test = 7 : 3$), whose grayscale similarity and inherent acquisition noise introduce significant interference during segmentation. Moreover, to evaluate segmentation performance in scenarios with weak or indistinct boundaries, we employed two endoscopic datasets, CVC-ClinicDB [51] and Kvasir [52] ($train : test = 7 : 3$), both well known for blurred contours and low-contrast lesion edges. This diverse dataset configuration ensures a thorough and reliable assessment of PLNet across varying levels of difficulty and imaging modalities.

4.2 | Experimental Setups

All experiments were conducted using the MMSegmentation framework on a single RTX 4090 GPU. During preprocessing, the raw images were resized to 512×512 and normalised, whereas the segmentation masks were binarised and resized to the same resolution. Random resizing and flipping were used for data augmentation. The model was optimised using AdamW with an initial learning rate of 6×10^{-5}, a weight decay of 0.01, and a Poly learning rate scheduler. The loss function was defined as a weighted sum of CrossEntropyLoss and DiceLoss, with a weight ratio of 1 : 2.

To evaluate the effectiveness and generalisation capability of the proposed model, we conducted systematic comparisons with a diverse set of representative medical image segmentation methods. The compared approaches include CNN-based models (U-Net, Attention U-Net, CMU-Net, and CMUNeXt), Mamba-based models (U-Mamba and Swin-UMamba), foundation model-based approaches (MedSAM and SAM-Adapter), and CNN–Transformer hybrid architectures (TransUNet, BEFUnet, and H2Former). This multi-paradigm comparison evaluates different modelling strategies, including convolutional local feature extraction, state-space-based contextual modelling, prompt-based segmentation, and CNN-Transformer hybrid representation learning.

For a fair and transparent comparison, all baseline methods except MedSAM were reproduced using publicly available implementations on our datasets under the same data splits, preprocessing, training, and testing protocols. We followed the official settings without additional dataset-specific tuning. For MedSAM, we used the available pretrained model and generated bounding-box prompts from the lesion localization annotations. Each bounding box was expanded by a 20-pixel margin to provide sufficient contextual coverage and approximate manually drawn box prompts. The reproduction status of the baseline methods is further summarised in Table 2.

4.3 | Evaluation Metrics

For model evaluation, three commonly used metrics in medical image segmentation were adopted, including Dice score, Intersection-over-Union (IoU), and accuracy (ACC). Dice score measures the overlap between the predicted segmentation and the ground truth, with particular emphasis on balancing false positives and false negatives. IoU evaluates the ratio between the intersection and the union of the predicted and ground-truth masks,

TABLE 2 | Reproduction status of the baseline methods used for comparison.

Category	Methods	Reproduction status
CNN-based	U-Net [5]	Reproduced with public code
	Attention U-Net [53]	Reproduced with public code
	CMU-Net [54]	Reproduced with public code
	CMUNeXt [55]	Reproduced with public code
Mamba-based	U-Mamba [16]	Reproduced with public code
	Swin-UMamba [18]	Reproduced with public code
CNN-Transformer hybrid	TransUNet [12]	Reproduced with public code
	BEFUnet [32]	Reproduced with public code
	H2Former [56]	Reproduced with public code
Foundation model-based	SAM-Adapter [21]	Reproduced with public code
	MedSAM [20]	Pretrained model with box prompt

providing a stricter assessment of segmentation consistency. ACC reflects the overall correctness of pixel-wise predictions across the entire image. Together, these metrics offer a comprehensive and reliable evaluation of segmentation performance.

4.4 | Results and Analysis

We compared PLNet with several representative segmentation models. As shown in Figure 6, PLNet and CMUNeXt exhibit relatively low computational complexity and parameter counts among the compared methods. PLNet achieves higher IoU and Dice scores than CMUNeXt while maintaining a lightweight architecture, demonstrating a favourable balance between segmentation performance and model complexity.

According to Table 3, PLNet achieves Dice scores of 88.81% and 88.62% on the ISIC2017 and ISIC2018 datasets, respectively, with only 21.47M parameters and 26.15G FLOPs. Compared with U-Mamba, PLNet obtains higher Dice scores on these two datasets while reducing FLOPs from 506.27G to 26.15G. These results suggest that PLNet provides a favourable balance between segmentation performance, model compactness, and computational complexity. Rather than simply pursuing the best result on every dataset, PLNet demonstrates competitive multi-dataset performance under relatively low FLOPs and model size, highlighting its potential as a lightweight medical image segmentation model.

We further evaluated PLNet on multiple medical image datasets. As reported in Table 4, PLNet achieves an IoU of 61.51% and a Dice score of 73.47% on the BUSI ultrasound dataset, indicating

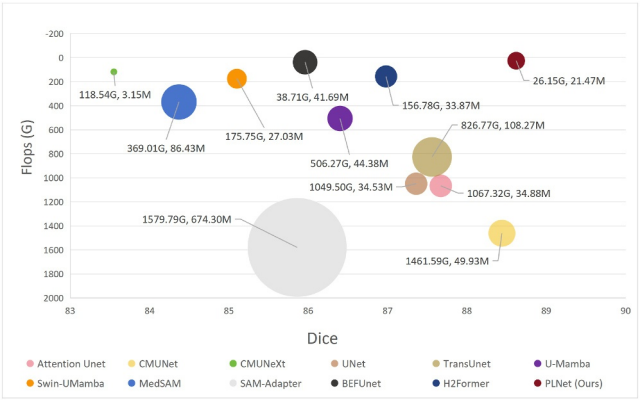


FIGURE 6 | Comparison of segmentation performance, FLOPs, and model size across different models on the ISIC2018 dataset, where circle size reflects the model parameter count.

competitive segmentation capability under challenging imaging conditions. On the CVC-ClinicDB and Kvasir colonoscopy datasets, PLNet also maintains competitive performance compared with representative methods. Although PLNet does not achieve the best result for every metric on every dataset, it provides a favourable balance between segmentation accuracy and computational complexity. In particular, compared with larger models such as MedSAM and SAM-Adapter, PLNet achieves comparable or better performance in several cases while requiring substantially lower model complexity.

As shown in Figure 7, the BUSI dataset is particularly challenging because lesion regions often occupy only a small proportion of the image, and their boundaries are usually difficult to distinguish from the surrounding tissue. Despite these difficulties, PLNet still produces relatively reliable segmentation results. Qualitative comparisons show that Swin-UMamba and MedSAM may include redundant regions outside the lesion in some ultrasound predictions, whereas BEFUnet tends to produce less distinct lesion boundaries. In contrast, PLNet generates masks that are visually closer to the ground truth in these examples. This suggests that the proposed encoder with multiple branches and the MoE fusion mechanism help capture boundary-related features in challenging ultrasound segmentation cases.

We designed three variants of the PLNet model: PLNet-Tiny, PLNet-Base, and PLNet-Large. The smallest variant, PLNet-Tiny, requires 14.42 GFLOPs and contains 6.18M parameters, whereas the largest variant, PLNet-Large, requires 121.49 GFLOPs and contains 36.86M parameters. This design offers flexibility for practical applications, enabling the selection of an appropriate model tailored to specific requirements. As illustrated in Table 5, model size has the most significant impact on colonoscopy datasets. Specifically, PLNet-Large achieves Dice score improvements of 9.59 and 10.01% points over PLNet-Tiny on the CVC-ClinicDB and Kvasir datasets, respectively. Across the remaining datasets, PLNet-Large generally achieves better performance than the smaller variants, indicating that increased model capacity is particularly beneficial for the two colonoscopy datasets.

TABLE 3 | Quantitative comparison of PLNet with competing methods on the ISIC2017 and ISIC2018 datasets.

Method	Params	FLOPs	ISIC2017/ISIC2018		
			IoU	Dice	ACC
Attention U-Net [53]	34.88M	1067.32G	76.36/78.68	86.21/87.67	81.13/84.62
U-Net [5]	34.53M	1049.5G	75.91/78.25	85.87/87.36	80.08/84.07
CMU-Net [54]	49.93M	1461.59G	78.19/79.41	87.13/88.44	82.54/86.19
CMUNeXt [55]	**3.15M**	118.54G	78.98/74.51	87.74/83.55	85.65/81.08
TransUNet [12]	108.27M	826.77G	79.84/78.77	88.23/87.56	88.75/88.04
U-Mamba [16]	44.38M	506.27G	78.28/78.99	86.21/86.40	**94.75/93.81**
Swin-UMamba [18]	55.06M	175.75G	74.49/77.19	83.09/85.10	94.12/93.80
MedSAM [20]	86.43M	369.01G	73.92/74.01	84.36/84.37	92.31/90.71
SAM-Adapter [21]	674.30M	1579.79G	75.96/77.39	84.51/85.86	94.28/91.66
BEFUnet [32]	41.69M	38.71G	**80.48**/78.49	87.66/85.96	92.05/90.56
H2Former [56]	33.87M	156.78G	79.83/78.93	87.32/86.98	92.89/90.75
PLNet-Base (ours)	21.47M	**26.15G**	79.87/**79.57**	**88.81/88.62**	88.04/89.53

Note: Best and second-best performances are indicated by **boldface** and underlining, respectively.

TABLE 4 | Quantitative comparison of PLNet and competing methods on the BUSI, CVC-ClinicDB, and Kvasir datasets.

Category	Method	BUSI			CVC			Kvasir		
		IoU	Dice	ACC	IoU	Dice	ACC	IoU	Dice	ACC
CNN	Attention U-Net [53]	53.02	68.09	64.11	78.33	87.17	88.99	73.05	82.79	80.08
	U-Net [5]	51.18	66.07	61.34	78.19	86.95	85.38	72.81	82.46	79.67
	CMU-Net [54]	60.62	72.40	71.01	79.20	87.67	85.44	75.87	85.21	81.90
	CMUNeXt [55]	57.21	71.11	66.86	79.31	87.73	88.15	75.19	84.33	79.53
Mamba	U-Mamba [16]	60.55	73.15	90.23	79.10	87.63	90.05	73.91	85.23	84.12
	Swin-UMamba [18]	60.43	72.92	89.96	79.04	87.53	89.72	74.46	85.52	84.23
SAM	MedSAM [20]	58.01	72.32	91.37	65.77	75.81	92.01	62.45	72.38	**85.67**
	SAM-Adapter [21]	58.18	71.48	91.83	63.58	73.49	91.46	73.49	75.82	83.46
Hybrid models	TransUNet [12]	60.59	72.94	75.21	**79.70**	87.75	**92.93**	76.81	85.53	84.69
	BEFUnet [32]	**63.24**	70.74	92.21	79.28	88.32	91.02	71.19	80.13	82.81
	H2Former [56]	53.74	64.78	**92.38**	78.63	86.14	92.03	**79.71**	86.15	84.79
	PLNet-Base (ours)	61.51	**73.47**	90.40	79.12	**88.34**	89.65	75.85	**86.27**	84.36

Note: Best and second-best performances are indicated by **boldface** and underlining, respectively.

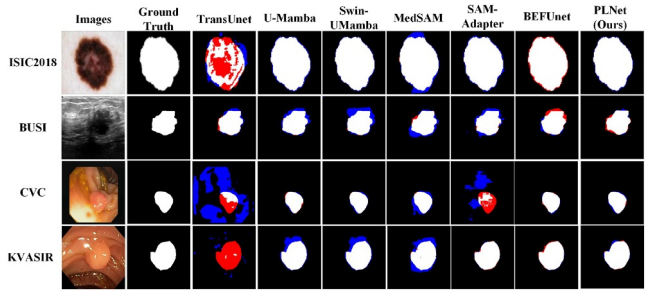


FIGURE 7 | Visualisation results of our proposed PLNet and other methods. White denotes overlap, red shows missing, and blue indicates redundant regions.

4.5 | Ablation Results

We performed ablation experiments on the ISIC2018 dataset to systematically evaluate the contribution of each component under a consistent experimental setting. Different encoder combinations were evaluated to investigate their effects on segmentation performance. As reported in Table 6, under the single encoder setting, the ViT encoder achieved the best results, with an IoU of 71.72% and a Dice score of 83.53%. Compared with the CNN and VSSM encoders, the ViT encoder shows stronger global semantic modelling capability through token interaction, which enables it to capture high-level contextual information. However, this stronger representation ability is accompanied by higher computational complexity.

Although the ViT encoder models global semantic relationships through token interaction, a single branch cannot simultaneously preserve fine-grained boundary cues and capture directional spatial dependencies. The ViT and VSSM branches are not redundant because they model contextual information through different mechanisms. Specifically, ViT captures global semantic relationships through token interaction, whereas VSSM propagates information along predefined spatial scanning paths to model directional dependencies and irregular lesion structures.

TABLE 5 | Performance comparison of different PLNet variants on five medical image segmentation datasets, covering model capacities from PLNet-Tiny (14.42 GFLOPs, 6.18M parameters) and PLNet-Base (26.15 GFLOPs, 21.47M parameters) to PLNet-Large (121.49 GFLOPs, 36.86M parameters).

	PLNet-tiny			PLNet-base			PLNet-large		
Dataset	**IoU**	**Dice**	**ACC**	**IoU**	**Dice**	**ACC**	**IoU**	**Dice**	**ACC**
ISIC2017	79.56	87.48	87.89	79.87	88.81	88.04	80.87	89.04	88.87
ISIC2018	78.18	87.28	87.79	79.57	88.62	89.53	80.89	89.43	87.34
BUSI	56.63	72.32	89.01	61.51	73.47	90.40	62.02	74.10	90.64
CVC	72.90	84.33	84.36	79.12	88.34	89.65	88.54	93.92	93.38
Kvasir	64.41	78.35	82.01	75.85	86.27	84.36	78.73	88.36	85.61

TABLE 6 | Ablation study on the ISIC2018 dataset evaluating the impact of different encoder combinations. CNN, ViT, and VSSM denote the convolutional encoder, Vision Transformer encoder, and VSSM encoder, respectively.

	Encoder					
Dataset	**CNN**	**ViT**	**VSSM**	**IoU**	**Dice**	**ACC**
ISIC2018	✓			66.24	79.69	86.95
		✓		71.72	83.53	90.43
			✓	64.84	78.66	**94.35**
	✓	✓		<u>72.67</u>	<u>84.17</u>	88.27
		✓	✓	72.15	83.82	<u>91.87</u>
	✓		✓	71.01	83.05	86.10
	✓	✓	✓	**79.57**	**88.62**	89.53

Note: Performance is reported in terms of IoU, Dice, and Accuracy. Best and second-best performances are indicated by **boldface** and underlining, respectively.

TABLE 7 | Ablation study on the ISIC2018 dataset evaluating the effect of different feature map fusion configurations in the pyramid-structured PLNet.

	Feature map						
Dataset	**1**	**2**	**3**	**4**	**IoU**	**Dice**	**ACC**
ISIC2018	✓				54.58	70.62	75.11
	✓	✓			66.80	80.10	80.79
	✓	✓	✓		67.51	80.60	<u>91.78</u>
				✓	<u>71.78</u>	<u>83.57</u>	**92.22**
	✓	✓	✓	✓	**79.57**	**88.62**	89.53

Note: Best and second-best performances are indicated by **boldface** and underlining, respectively.
Feature maps 1–4 denote multi-level representations used for fusion.

Their combination therefore provides complementary semantic and spatial representations.

Among the dual-encoder configurations, the CNN–ViT combination achieved the best performance, with an IoU of 72.67% and a Dice score of 84.17%. The CNN branch preserves local texture and boundary cues, while the ViT branch models global semantic context. However, this configuration does not explicitly model directional spatial dependencies. After introducing the VSSM branch, the IoU and Dice scores increased by 6.90 and 4.45% points, respectively. These improvements indicate that the directional state-space modelling of VSSM complements the local representation of CNN and the global semantic modelling of ViT.

We also investigated the effect of feature map fusion on the performance of PLNet. Since PLNet adopts a pyramid architecture, we analysed the contributions of feature maps from different depths. As reported in Table 7, using only the shallowest feature map achieved an IoU of 54.58% and a Dice score of 70.62%. This indicates that shallow features preserve local texture and boundary details, but lack sufficient semantic information. As deeper feature maps were gradually incorporated, the segmentation performance consistently improved.

We further tested the setting using only the deepest feature map, which achieved an IoU of 71.78% and a Dice score of 83.57%. Although deep features contain stronger semantic context, they may lose spatial details during downsampling, which limits boundary accuracy. In contrast, fusing all feature maps achieved the best performance, with an IoU of 79.57% and a Dice score of 88.62%. This is particularly important for medical image segmentation, where lesions are often small, low contrast, and have unclear boundaries. Therefore, multi-level feature fusion helps combine shallow boundary information with deep semantic discrimination, leading to more accurate mask prediction.

We further compared different encoder feature fusion strategies and observed clear performance differences among the four fusion methods. As reported in Table 8, the MoE-based fusion strategy achieved the best performance, with an IoU of 79.57% and a Dice score of 88.62%, surpassing the other fusion methods. This advantage mainly stems from the ability of MoE to select more informative features while reducing the influence of redundant or less useful ones, thereby enhancing the segmentation capability of the model. By contrast, weighted fusion produced the worst results, with an IoU of 69.42% and a Dice score of 81.95%, which were 10.15 and 6.67% points lower than those of the MoE-based strategy, respectively. In addition, under the additive fusion setting, combining multi-level feature maps improved the IoU and Dice scores by 3.83 and 3.89% points, respectively, compared with using only the lowest-level feature map. These results indicate that MoE fusion is most effective for cross-encoder feature aggregation in PLNet.

We conducted experiments to evaluate different scanning directions and fusion strategies in the SS2D module. As shown in Table 9, increasing the number of scanning directions from

TABLE 8 | Comparison of different feature fusion strategies for PLNet on the ISIC2018 dataset.

Dataset	Feature fusion method (encoder)	IoU	Dice	ACC
ISIC2018	Weighted fusion	69.42	81.95	84.26
	Additive fusion (last layer)	71.78	83.57	**92.22**
	Additive fusion	75.61	87.46	88.99
	MoE fusion	**79.57**	**88.62**	89.53

Note: Best and second-best performances are indicated by **boldface** and underlining.

TABLE 9 | Effect of different scanning directions and feature fusion strategies on the segmentation performance of PLNet on the ISIC2018 dataset.

Dataset	Scanning direction	Directional fusion method in the SS2D module	IoU	Dice	ACC
ISIC2018	4 ways	Weighted fusion	72.10	83.79	**94.08**
		Additive fusion	74.12	85.14	93.06
		MoE fusion	71.94	83.68	89.28
	8 ways	Weighted fusion	72.71	84.20	92.91
		Additive fusion	**79.57**	**88.62**	89.53
		MoE fusion	67.20	80.50	93.57

Note: Best and second-best performances are indicated by **boldface** and underlining, respectively.

four to eight improved segmentation performance under the weighted and additive fusion settings, whereas no corresponding improvement was observed with MoE fusion. In particular, under additive fusion, the IoU and Dice scores increased by 5.45 and 3.48% points, respectively. This improvement indicates that the additional diagonal scanning paths provide complementary boundary cues and directional spatial dependencies that are not fully captured by the four axial directions. The best performance was achieved by combining eight scanning directions with additive fusion, yielding an IoU of 79.57% and a Dice score of 88.62%.

This result may be attributed to the characteristics of directional feature aggregation within SS2D. Since features from different scanning directions are generated from the same input feature map, each direction preserves distinct but complementary responses associated with lesion boundaries, local structures, and spatial dependencies. Additive fusion retains the responses from all scanning paths and therefore facilitates their joint aggregation. In contrast, sparse MoE fusion activates only a subset of experts and may assign relatively small or zero routing weights to weaker yet useful directional responses, thereby limiting the contribution of complementary scanning paths in this setting.

This behaviour differs from the fusion of features from different encoders, where the CNN, ViT, and VSSM branches produce heterogeneous representations with different semantic characteristics and possible redundancy. In this case, the MoE mechanism is beneficial because its expert selection process can adaptively emphasise informative representations and reduce redundant responses. Therefore, additive fusion is more suitable for directional feature aggregation within SS2D, whereas MoE is more appropriate for integrating heterogeneous features from different encoders. The best-performing settings in Tables 8 and 9 correspond to the same final PLNet configuration, where MoE is used for feature fusion among different encoders and additive fusion is used for directional aggregation within the 8-direction SS2D module. Therefore, the same final performance is reported in both tables from different ablation perspectives.

TABLE 10 | Analysis of expert utilization in the MoE-based cross-encoder fusion module.

Stage	Dominant expert	Max Top-1 ratio	Avg. Active experts	Entropy	Effective experts
Stage 0	Expert 1	1.0000	1.0145	0.3454	1.6823
Stage 1	Expert 3	0.5636	1.9091	0.8288	3.1821
Stage 2	Expert 3	0.5418	1.4582	0.6507	2.5924
Stage 3	Expert 2	0.9091	1.3855	0.5303	2.1356

To further support this design choice, we additionally analysed the expert utilization of the MoE module on the test set. The gating weights of each MoE layer were recorded, and the Top-1 expert ratio, average active experts, normalised gate entropy, and effective number of experts were computed. As shown in Table 10, the MoE module exhibits stage-specific expert preferences. Although some stages show stronger dominant-expert tendencies, the selected dominant experts differ across stages, indicating that the router does not collapse to the same expert globally. Instead, it learns stage-dependent routing patterns for different feature levels. This observation further supports the use of MoE for cross-encoder feature fusion, where heterogeneous representations from CNN, ViT, and VSSM branches require adaptive and selective integration.

5 | Conclusions

PLNet is a pyramid-based medical image segmentation model designed to balance segmentation accuracy, model size, and computational complexity. Its core design is a collaborative three-branch encoder combined with adaptive MoE-based feature fusion and a lightweight decoder. The CNN branch captures local texture and boundary details, the ViT branch models global semantic context through token interaction, and the VSSM branch captures structured spatial dependencies through multi-directional scanning. The MoE fusion mechanism emphasises informative representations and reduces redundant responses, while the lightweight decoder provides a compact prediction head with low computational cost. Extensive experiments demonstrate that PLNet achieves competitive segmentation performance across multiple benchmark datasets while maintaining low computational complexity, showing a favourable balance between segmentation accuracy and computational complexity.

Despite its encouraging results, PLNet still faces limitations in clinical scenarios with limited annotated data, where data scarcity may restrict model generalisation, as well as in resource-constrained deployment environments. Future work will therefore focus on improving data efficiency under limited supervision, enhancing robustness across different imaging conditions, and

extending the proposed framework to a broader range of medical imaging tasks.

Funding

This work was supported by the Jiangsu Province Frontier Technology R&D Program; (Grant BF2025613), the Clinical Competence Improvement Project of Jiangsu Province Hospital (Grant JSPH-MB-2022-1), Jiangsu Provincial Key Research and Development Program (Grant BE2023818) and National Natural Science Foundation of China (Grant 12227808).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are publicly available from the corresponding benchmark repositories, including ISIC2017, ISIC2018, BUSI, CVC-ClinicDB, and Kvasir. The processed data and implementation details used in this study are available from the corresponding author upon reasonable request. The source code is publicly available at the PLNet repository. The source code is available at https://github.com/lzw514601882/PLNet.

References

1. Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, (Springer, 2018), 3–11, https://doi.org/10.1007/978-3-030-00889-5_1.

2. A. Hatamizadeh, Y. Tang, V. Nath, et al., "UNETR: Transformers for 3D Medical Image Segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (IEEE, 2022), 574–584, https://doi.org/10.1109/WACV51458.2022.00181.

3. A. Kurz, K. Hauser, H. A. Mehrtens, et al., "Uncertainty Estimation in Medical Image Classification: Systematic Review," *JMIR Medical Informatics* 10, no. 8 (2022): e36427, https://doi.org/10.2196/36427.

4. C. Yin, H. Su, Y. Xie, et al., "Estimating in Vivo Power Deposition Density in Thermotherapies Based on Ultrasound Thermal Strain Imaging," *Journal of the Acoustical Society of America* 157, no. 1 (2025): 606–617, https://doi.org/10.1121/10.0034880.

5. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Lecture Notes in Computer Science*, Vol. 9351 (Springer, 2015), 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

6. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, K. H. Maier-Hein, "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation," *Nature Methods* 18, no. 2 (2021): 203–211, https://doi.org/10.1038/s41592-020-01008-z.

7. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (IEEE, 2016), 565–571, https://doi.org/10.1109/3DV.2016.79.

8. Z. Gu, J. Cheng, H. Fu, et al., "Ce-Net: Context Encoder Network for 2d Medical Image Segmentation," *IEEE Transactions on Medical Imaging* 38, no. 10 (2019): 2289–2297, https://doi.org/10.1109/tmi.2019.2903562.

9. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the 9th International Conference on Learning Representations (ICLR)* (OpenReview.net, 2021).

10. H. Cao, Y. Wang, J. Chen, et al., "Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation," in *Computer Vision – ECCV 2022 Workshops* (Springer Nature Switzerland, 2022), 205–218.

11. X. Liu, P. Gao, T. Yu, F. Wang, and R.-Y. Yuan, "Cswin-Unet: Transformer U-Net With Cross-Shaped Windows for Medical Image Segmentation," *Information Fusion* 113 (2025): 102634, https://doi.org/10.1016/j.inffus.2024.102634.

12. J. Chen, J. Mei, X. Li, et al., "Transunet: Rethinking the U-Net Architecture Design for Medical Image Segmentation Through the Lens of Transformers," *Medical Image Analysis* 97 (2024): 103280, https://doi.org/10.1016/j.media.2024.103280.

13. G. Xu, X. Wu, X. Zhang, and X. He, "Levit-Unet: Make Faster Encoders With Transformer for Medical Image Segmentation," in *Pattern Recognition and Computer Vision-PRCV 2023*, 14432 (Springer Nature Singapore, 2023), 42–53, https://doi.org/10.1007/978-981-99-8543-2_4.

14. A. Gu, K. Goel, and C. Ré, "Efficiently Modeling Long Sequences With Structured State Spaces," in *International Conference on Learning Representations*, (2022).

15. Y. Liu, Y. Tian, Y. Zhao, et al., "VMamba: Visual State Space Model," in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)* (Neural Information Processing Systems Foundation, Inc., 2024), 103031–103063, https://doi.org/10.52202/079017-3273.

16. J. Ma, F. Li, and B. Wang, "U-Mamba: Enhancing Long-Range Dependency for Biomedical Image Segmentation," arXiv preprint arXiv:2401.04722, 2024, https://doi.org/10.48550/arXiv.2401.04722.

17. Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "Segmamba: Long-Range Sequential Modeling Mamba for 3d Medical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2024*, 15008 (Springer Nature Switzerland, 2024), 578–588, https://doi.org/10.1007/978-3-031-72111-3_54.

18. J. Liu, H. Yang, H.-Y. Zhou, et al., "Swin-UMamba: Mamba-Based Unet With Imagenet-Based Pretraining," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2024*, 15009 (Springer Nature Switzerland, 2024), 615–625, https://doi.org/10.1007/978-3-031-72114-4_59.

19. O. Siméoni, H. V. Vo, M. Seitzer, et al., "DINOv3," *Transactions on Machine Learning Research* (2026).

20. J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment Anything in Medical Images," *Nature Communications* 15, no. 1 (2024): 654, https://doi.org/10.1038/s41467-024-44824-z.

21. T. Chen, L. Zhu, C. Ding, et al., "Sam-Adapter: Adapting Segment Anything in Underperformed Scenes," in *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (IEEE, 2023), 3359–3367, https://doi.org/10.1109/ICCVW60793.2023.00361.

22. T. Chen, A. Lu, L. Zhu, et al., "SAM2-Adapter: Evaluating & Adapting Segment Anything 2 in Downstream Tasks: Camouflage, Shadow, Medical Image Segmentation, and More," in *ICLR 2025 Workshop on Foundation Models in the Wild* (OpenReview.net, 2025).

23. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2017), 2261–2269, https://doi.org/10.1109/CVPR.2017.243.

24. H. Huang, L. Lin, R. Tong, et al., "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2020), 1055–1059, https://doi.org/10.1109/ICASSP40776.2020.9053405.

25. Y. Peng, D. Z. Chen, and M. Sonka, "U-Net V2: Rethinking the Skip Connections of U-Net for Medical Image Segmentation," in *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)* (IEEE, 2025), 1–5, https://doi.org/10.1109/ISBI60581.2025.10980742.

26. Z. Luo, X. Zhu, L. Zhang, and B. Sun, "Rethinking U-Net: Task-Adaptive Mixture of Skip Connections for Enhanced Medical Image Segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, no. 6 (AAAI Press, 2025), 5874–5882, https://doi.org/10.1609/aaai.v39i6.32627.

27. Z. Tang, Y. Fu, M. Li, H. Liang, Y. Tang, and J. Wen, "MPR-Net: Medicinal Plant Recognition Network With Dual-Branch Attention Fusion," *Pattern Recognition* 171 (2026): 112185, https://doi.org/10.1016/j.patcog.2025.112185.

28. C. Li, X. Liu, W. Li, et al., "U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, no. 5 (AAAI Press, 2025), 4652–4660, https://doi.org/10.1609/aaai.v39i5.32491.

29. H. Wang, Z. Zhao, Q. Liu, and S. Wang, "ResU-KAN: A Medical Image Segmentation Model Integrating Residual Convolutional Attention and Atrous Spatial Pyramid Pooling," *Applied Intelligence* 55, no. 7 (2025): 568, https://doi.org/10.1007/s10489-025-06467-5.

30. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training Data-Efficient Image Transformers & Distillation Through Attention," in *Proceedings of the 38th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, Vol. 139 (PMLR, 2021), 10347–10357.

31. S. Wang, G. Li, M. Gao, et al., "GH-UNet: Group-Wise Hybrid Convolution-vit for Robust Medical Image Segmentation," *npj Digital Medicine* 8, no. 1 (2025): 426, https://doi.org/10.1038/s41746-025-01829-2.

32. O. N. Manzari, J. M. Kaleybar, H. Saadat, and S. Maleki, "BEFUnet: A Hybrid CNN-Transformer Architecture for Precise Medical Image Segmentation," arXiv preprint arXiv:2402.08793, 2024, https://doi.org/10.48550/arXiv.2402.08793.

33. J. Wen, J. Long, X. Lu, C. Liu, X. Fang, and Y. Xu, "Partial Multi-view Incomplete Multilabel Learning via Uncertainty-Driven Reliable Dynamic Fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

34. A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling With Selective State Spaces," in *First Conference on Language Modeling (COLM)* (OpenReview.net, 2024).

35. W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative Global-Local Networks for Memory-Efficient Segmentation of Ultra-High Resolution Images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2019), 8924–8933, https://doi.org/10.1109/CVPR.2019.00913.

36. H. K. Cheng, J. Chung, Y.-W. Tai, and C.-K. Tang, "CascadePSP: Toward Class-Agnostic and Very High-Resolution Segmentation via Global and Local Refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2020), 8887–8896, https://doi.org/10.1109/CVPR42600.2020.00891.

37. B. Chen, X. Zou, Y. Zhang, et al., "Leformer: A Hybrid CNN-Transformer Architecture for Accurate Lake Extraction From Remote Sensing Imagery," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, (2024), 1–5.

38. H. Cheng, X. Liu, J. Zhang, et al., "GLMKD: Joint Global and Local Mutual Knowledge Distillation for Weakly Supervised Lesion Segmentation in Histopathology Images," *Expert Systems With Applications* 279 (2025): 127425, https://doi.org/10.1016/j.eswa.2025.127425.

39. W. Zhao, L. Zhang, Y. Zhao, X. Bai, Z. Xu, and Y. Li, "GLAD-Net: Global-Local Adaptive Fusion and Cross-Stage Distillation for Cross-Level Multi-Scale Medical Image Segmentation," in *Proceedings of the 21st International Conference on Intelligent Computing (ICIC 2025)* (2025), 2083–2097, Poster Volume II.

40. F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2017), 1800–1807, https://doi.org/10.1109/CVPR.2017.195.

41. S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation," in *Computer Vision—ECCV 2018, Lecture Notes in Computer Science*, Vol. 11214 (Springer, 2018), 561–580, https://doi.org/10.1007/978-3-030-01249-6_34.

42. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Computer Vision—ECCV 2018, Lecture Notes in Computer Science*, Vol. 11211 (Springer, 2018), 3–19, https://doi.org/10.1007/978-3-030-01234-2_1.

43. W. Wang, E. Xie, X. Li, et al., "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021), 568–578, https://doi.org/10.1109/ICCV48922.2021.00061.

44. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation With Transformers," *Advances in Neural Information Processing Systems 34* (Neural Information Processing Systems Foundation, Inc., 2021), 12077–12090.

45. W. Yu, M. Luo, P. Zhou, et al., "MetaFormer Is Actually What You Need for Vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2022), 10819–10829, https://doi.org/10.1109/CVPR52688.2022.01055.

46. Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021), 10012–10022, https://doi.org/10.1109/ICCV48922.2021.00986.

47. N. C. F. Codella, D. Gutman, M. E. Celebi, et al., "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)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (IEEE, 2018), 168–172, https://doi.org/10.1109/ISBI.2018.8363547.

48. N. C. F. Codella, V. Rotemberg, P. Tschandl, et al., "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," (2019), https://doi.org/10.48550/arXiv.1902.03368.

49. P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions," *Scientific Data* 5, no. 1 (2018): 180161, https://doi.org/10.1038/sdata.2018.161.

50. W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of Breast Ultrasound Images," *Data in Brief* 28 (2020): 104863, https://doi.org/10.1016/j.dib.2019.104863.

51. J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA Maps for Accurate Polyp Highlighting in Colonoscopy: Validation vs. Saliency Maps From Physicians," *Computerized Medical Imaging and Graphics* 43 (2015): 99–111, https://doi.org/10.1016/j.compmedimag.2015.02.007.

52. D. Jha, P. H. Smedsrud, M. A. Riegler, et al., "Kvasir-SEG: A Segmented Polyp Dataset," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26* (Springer, 2020), 451–462.

53. O. Oktay, J. Schlemper, L. Le Folgoc, et al., "Attention U-Net: Learning Where to Look for the Pancreas," arXiv preprint arXiv:1804.03999, 2018, https://doi.org/10.48550/arXiv.1804.03999.

54. F. Tang, L. Wang, C. Ning, M. Xian, and J. Ding, "CMU-Net: A Strong ConvMixer-Based Medical Ultrasound Image Segmentation Network," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (IEEE, 2023), 1–5, https://doi.org/10.1109/ISBI53787.2023.10230609.

55. F. Tang, J. Ding, Q. Quan, L. Wang, C. Ning, and S. K. Zhou, "CMUNeXt: An Efficient Medical Image Segmentation Network Based on Large Kernel and Skip Fusion," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)* (IEEE, 2024), 1–5, https://doi.org/10.1109/ISBI56570.2024.10635609.

56. A. He, K. Wang, T. Li, C. Du, S. Xia, and H. Fu, "H2former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation," *IEEE Transactions on Medical Imaging* 42, no. 9 (2023): 2763–2775, https://doi.org/10.1109/tmi.2023.3264513.