MedLiteNet: Lightweight Hybrid Medical Image Segmentation Model for the Medical Internet of Things (MIoT)

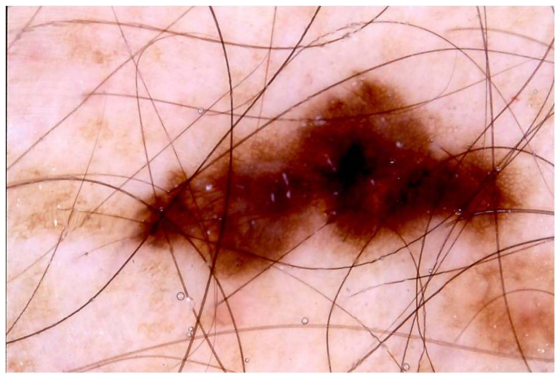
Pengyang Yu, Haoquan Wang, Gerard Marks, Tahar Kechadi, Laurence T. Yang, *Fellow, IEEE*, Sahraoui Dhelim, and Nyothiri Aung

***Abstract*—Accurate skin-lesion segmentation remains a key technical challenge for computer-aided diagnosis of skin cancer. Convolutional neural networks, while effective, are constrained by limited receptive fields and thus struggle to model long-range dependencies. Vision Transformers capture global context, yet their quadratic complexity and large parameter budgets hinder use on the small-sample medical datasets common in dermatology. We introduce the MedLiteNet, a lightweight CNN–Transformer hybrid tailored for dermoscopic segmentation that achieves high precision through hierarchical feature extraction and multiscale context aggregation. The encoder stacks depthwise mobile inverted bottleneck blocks to curb computation, inserts a bottleneck-level cross-scale token-mixing unit to exchange information between resolutions, and embeds a boundary-aware self-attention module to sharpen lesion contours. On the ISIC 2018 benchmark, a single MedLiteNet model attains a Dice score of 0.897 ± 0.010 and an IoU of 0.821 ± 0.015 with fewer than $3.3\,\mathrm{M}$ parameters. A performance-weighted ensemble of three complementary variants raises accuracy to 0.904 ± 0.012 Dice and 0.830 ± 0.018 IoU while keeping the total parameter count below $10\,\mathrm{M}$—over 90% smaller than Vision-Transformer backbones. Qualitative results confirm superiority on irregular borders, low-contrast regions and multiscale lesions, indicating MedLiteNet's suitability for real-time, resource-aware computer-aided dermatology.**

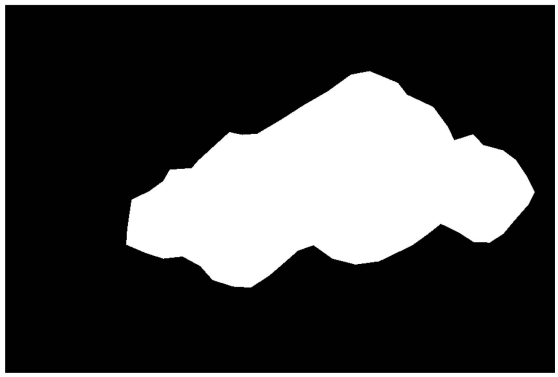
***Index Terms*—Attention mechanism, boundary detection, convolutional neural network, lightweight model, medical image segmentation, skin lesions, transformer.**

I. Introduction

SKIN cancer (especially melanoma) is a serious disease that threatens human health worldwide. Accurate automatic segmentation of skin lesion areas is of great significance for computer-aided diagnosis and treatment planning. It can help doctors more clearly assess the boundaries and areas of lesions, thereby improving the accuracy of early diagnosis and treatment effects.



(a) Original dermoscopic image



(b) Lesion segmentation result

Fig. 1. Comparison of original image and segmentation result for ISIC_0000214.

However, skin lesion segmentation faces many challenges: the color and texture of lesions in skin images vary widely, and often have low contrast with the surrounding normal skin (see Fig. 1); the shape and size of lesions vary significantly under different patients and imaging conditions, and the boundaries may be blurred. Traditional image segmentation algorithms (such as threshold methods, active contour models, etc.) require a lot of manual parameter adjustment and are difficult to cope with the complex and changeable morphology of skin lesions.

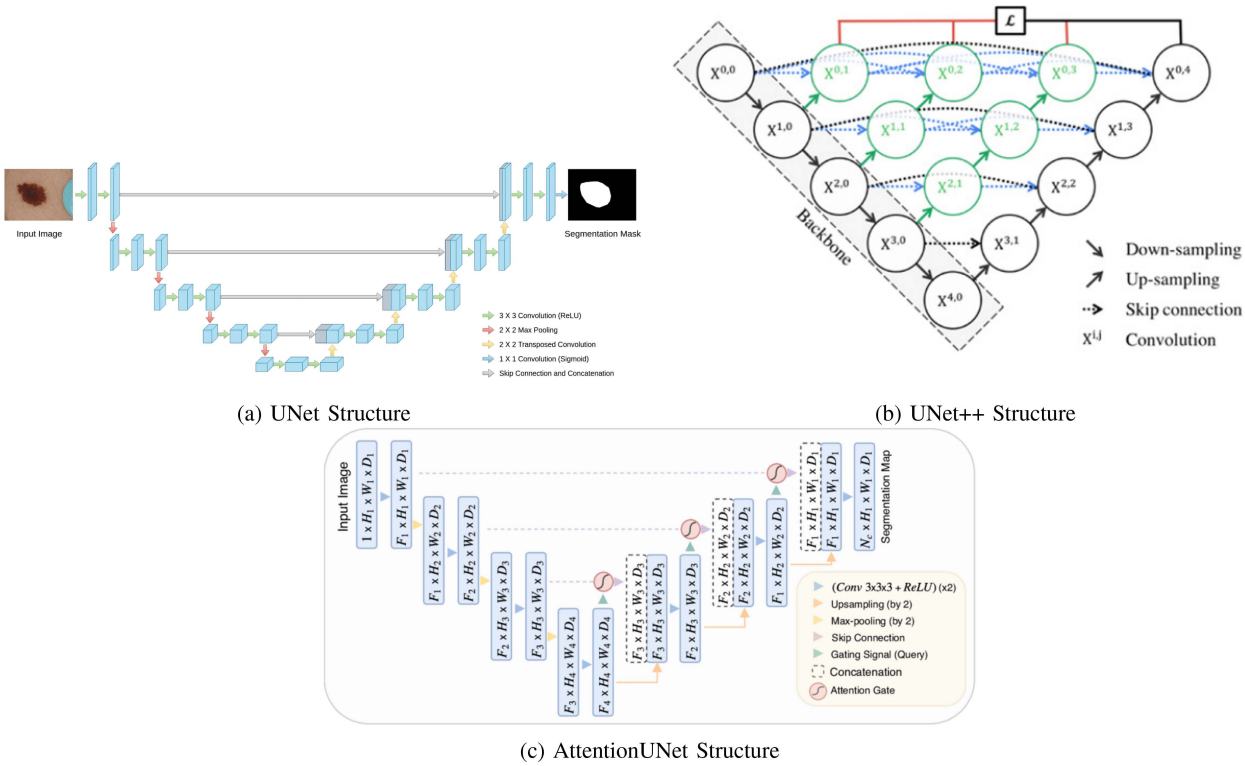
Received 10 November 2025; revised 28 November 2025; accepted 15 December 2025. Date of publication 25 December 2025; date of current version 9 February 2026. This work was supported by Open Access funding provided by University College Dublin. Recommended by Lead Guest Editor Xun Shao and Guest Editor Xun Shao. *(Corresponding author: Pengyang Yu.)*

Pengyang Yu, Tahar Kechadi, and Nyothiri Aung are with the School of Computer Science, University College Dublin, D04 YX26 Dublin, Ireland (e-mail: pengyang.yu@ucdconnect.ie).

Haoquan Wang, Gerard Marks, and Sahraoui Dhelim are with the School of Computing, Dublin City University, D09 Y5N0 Dublin, Ireland.

Laurence T. Yang is with the School of Computer and Artificial Intelligence, Zhengzhou University, 450001 Zhengzhou, China.

Digital Object Identifier 10.1109/JSAS.2025.3648685



(a) UNet Structure (b) UNet++ Structure

(c) AttentionUNet Structure

Fig. 2. Comparison of three common convolutional segmentation network structures.

Therefore, in recent years, researchers have turned their attention to data-driven deep learning methods, hoping to use their powerful feature learning capabilities to solve the problem of skin cancer segmentation.

A. Difficulties of Existing Technologies

Traditional threshold, region growing and active contour methods rely on fixed priors and are difficult to adapt to images with blurred lesion edges and uneven brightness under different patients, imaging devices, and lighting conditions. CNN, represented by U-Net [1], significantly improves segmentation accuracy by fusing multiscale features through an encoding–decoding structure, but the local receptive field is limited, and it is difficult to capture long-distance dependencies and large-scale deformations. Vision Transformer (ViT) uses self-attention to achieve global modeling, but lacks inductive biases such as translation invariance. It is easy to underfit in small sample medical scenarios, and the self-attention calculation grows quadratically with the resolution, the reasoning is time-consuming and the edge characterization is insufficient.

Although hybrid architectures in recent years (such as TransUNet, Swin-Unet, TransFuse [2], FAT-Net) have both local and global advantages, they mostly rely on heavy backbones such as ResNet-50 or ViT-B, and the number of parameters and computing power overhead are tens of megabytes, which is not conducive to mobile or real-time clinical deployment. Therefore, how to effectively compress the model size and reduce the reasoning delay while maintaining segmentation accuracy has become a key issue for medical image segmentation to move from the laboratory to real applications.

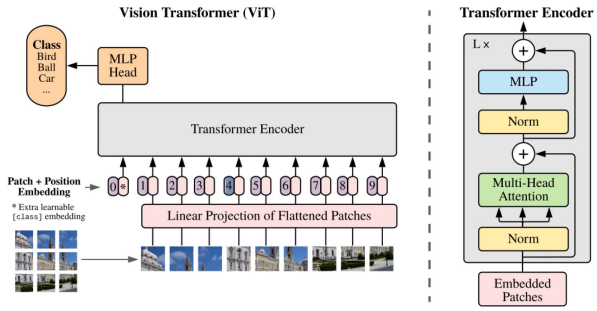


Fig. 3. ViT structure [12] for image-level representation learning via patch embeddings and self-attention.

B. Our Method and Contributions

MedLiteNet proposes a lightweight CNN-Transformer hybrid architecture for ISIC 2018 dermoscopy images, with MobileNetV2 inverted residual as the backbone. The cross-scale local–global block fuses the local texture of the depthwise separable convolution and the global dependencies captured by the self-attention mechanism, and injects boundary-aware weights into the attention mechanism to enhance blurred contours.

The decoder uses spatial and channel squeeze-and-excitation (SCSE) [3] recalibration and enhanced ASPP in parallel, and multiscale aggregation maintains good localization accuracy while maintaining fine localization. The model has only about 3 million parameters. After size-increment training optimization, the single-frame inference time for 256 × 256 images is about 1 ms (RTX A6000), with extremely low computational and storage overhead.

In the ISIC 2018 evaluation, the final integrated test achieved a validation set Dice of 0.913/IoU 0.84 and a test set Dice of 0.905/0.83, achieving lightweight while maintaining segmentation performance that is almost the same as the best method.

The experiment shows that the carefully designed lightweight hybrid architecture can maintain high-precision skin lesion segmentation at extremely low computational complexity, meeting the deployment requirements of resource-constrained scenarios (such as mobile devices and primary healthcare), and is expected to promote the popularization of early skin cancer screening technology.

Our contributions are summarized as follows.

1) We propose a "local–global–boundary" triple-coupled architecture that delivers clinical-grade accuracy with a mobile-scale footprint.
2) We developed a boundary-aware self-attention mechanism that explicitly reinforces contour representation.
3) We propose a progressive size-increment training and FP16 inference recipe that processes a 256 × 256 image in 1 ms raw latency (23 ms with sixfold TTA) [4].
4) Extensive experiments showing that carefully designed lightweight networks, even when ensembled, can rival heavy models in realistic dermatology scenarios.

II. Related Works

A. CNN-Based Segmentation Methods

CNN has become the mainstream technology for medical image segmentation with its excellent local feature extraction capabilities [5]. Among them, the U-Net model with an encoder–decoder structure is the most representative. U-Net extracts high-level semantic features by downsampling and fuses high-resolution details through skip connections, achieving good segmentation of fine structures in biomedical images. Various improved models of U-Net have also emerged. For example, the U-Net++ model introduces dense skip connections between encoder layers, further optimizing feature fusion through nested structures; Attention U-Net [6], [7], R2U-Net [8] introduces an attention gate mechanism at the skip connection to suppress irrelevant regional features to highlight the target area. A structural comparison of these common convolutional models is shown in Fig. 2. These CNN models have achieved excellent results on many public datasets.

However, the inherent limitation of CNN is that the receptive field is limited and it mainly focuses on local neighborhood pixels. Although the global field of view can be expanded by deepening the network or dilated convolution, it is still difficult to effectively capture long-distance dependencies. In addition, CNN usually relies on downsampling for feature compression, which may lead to the loss of fine-grained spatial information [9], [10]. For skin lesions with blurred boundaries and small sizes, pure CNN models are often unable to cope with it.

B. Transformer-Based Segmentation Methods

The introduction of the Transformer architecture [11] provides a new opportunity for global information modeling. Transformer models the correlation between any two positions in a sequence through the self-attention mechanism, and has made breakthroughs in tasks such as natural image classification. Vision Transformer (ViT) [12], whose architecture is illustrated in Fig. 3, divides images into patch embedding sequences, and for the first time proves the feasibility of pure Transformer in visual tasks [13].

In the field of medical image segmentation, the TransUNet model proposed by Chen et al. [14] embeds the pretrained ViT into the U-Net encoder, uses the Transformer to encode long-range dependencies, and uses the features extracted by CNN to retain local details. TransUNet significantly improves the accuracy in tasks such as multiorgan segmentation. Subsequently, models such as Swin-Unet [15] use a hierarchical window attention mechanism to improve the efficiency of the Transformer, which can reduce computational overhead while maintaining global modeling.

Although the Transformer has expanded the receptive field, it still has some problems in the application of medical image segmentation: on the one hand, the Transformer model has a huge number of parameters and requires massive data pretraining. The scale of medical datasets such as ISIC is relatively limited, and direct training of the Transformer is prone to overfitting [16]; on the other hand, forcibly embedding the Transformer module into the CNN architecture may bring about differences in feature distribution, resulting in inconsistent fusion, thereby affecting segmentation performance. In addition, ViT's method of dividing images into patches will lose some spatial details, which is also a major obstacle for segmentation tasks that require pixel-level fine positioning.

C. CNN-Transformer Hybrid Architecture

To reconcile the strength of CNNs in encoding local details with the Transformer's ability to capture long-range context, a growing number of studies have adopted hybrid designs. Early attempts, such as TransUNet [14], grafted a pretrained ViT encoder onto a U-Net backbone, achieving 88.5% Dice / 83.7% IoU at the cost of 105 M parameters.

Subsequent efforts reduced complexity: FAT-Net (2022) combined a lightweight CNN with a windowed self-attention mechanism, cutting parameters to 28 M while slightly improving the Dice to 89.0%. More recent methods have further pushed the efficiency boundary—BACANet [17] introduced boundary-aware convolutions and attention gating, reaching a state-of-the-art 92.1% Dice / 85.4% IoU with only 7.56 M parameters. However, even this footprint may be prohibitive for mobile or point-of-care devices.

Our proposed *MedLiteNet* (2025) compresses the hybrid paradigm to just 3.2 M parameters—more than 98% less than TransUNet—while maintaining competitive accuracy (Dice 90.5%, IoU 83.0%). The comparison in Table I shows that model size has fallen by two orders of magnitude in four years, yet a truly "Pareto-optimal" balance between accuracy and efficiency remains elusive.

MedLiteNet aims to bridge this gap through the following.

1) Depth-based MBConv encoder—leverages depthwise separable mobile inverted bottleneck (MBConv) blocks to

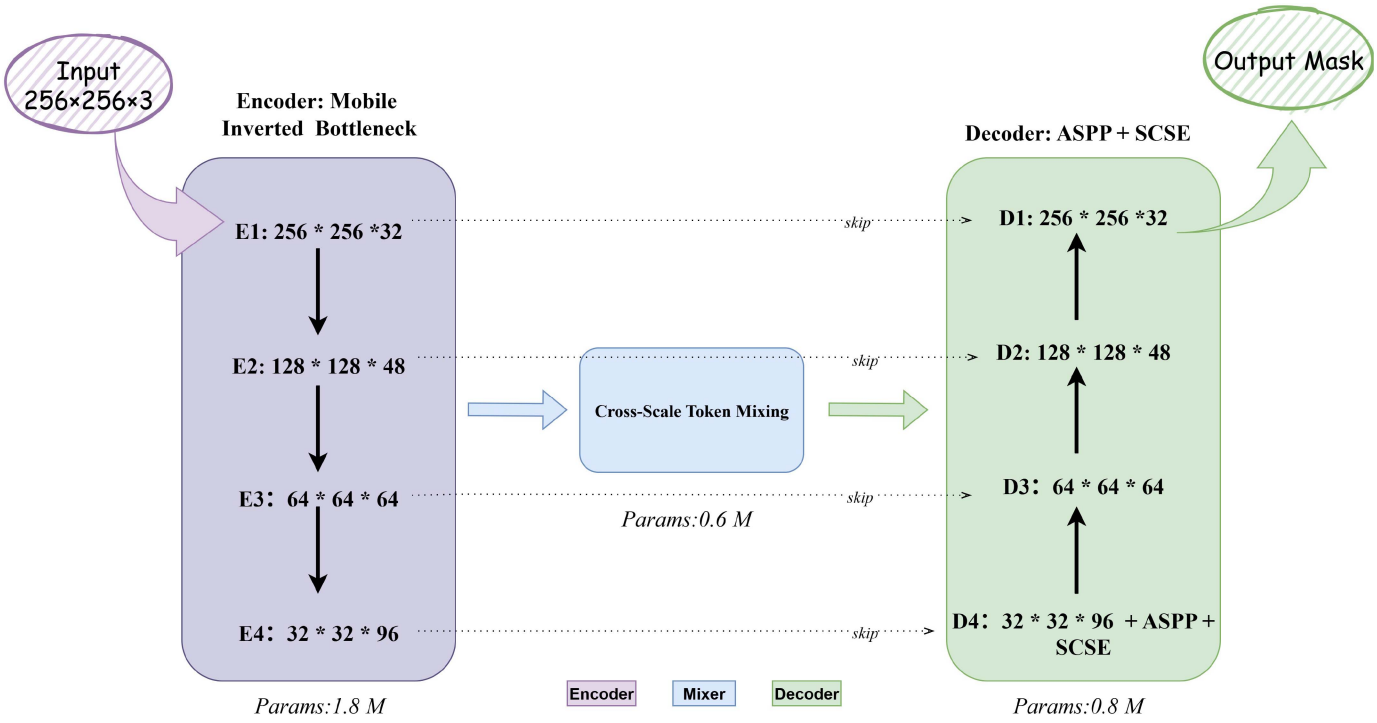


Fig. 4. Overall architecture of the proposed MedLiteNet, which consists of a lightweight encoder, a cross-scale token mixer, and a decoder equipped with ASPP and SCSE modules.

TABLE I
SUMMARY OF CNN–TRANSFORMER HYBRID MODELS

Model	Year	Params	Dice (%)	IoU (%)
TransUNet [10]	2021	105M	88.5	83.7
FAT-Net [8]	2022	28M	89.0	80.2
BACANet [3]	2024	7.56M	92.1	85.4
MedLiteNet	2025	3.2M	90.5	83.0

build a compact backbone while preserving local texture representation.

2) Cross-scale token hybrid module—fuses convolutional features with cross-scale tokens, using self-attention to capture long-range context.
3) ASPP + SCSE decoder—combines atrous spatial pyramid pooling (ASPP) with parallel SCSE attention for multi-scale semantics and fine-grained boundary recovery.

This lightweight yet high-performance configuration makes *MedLiteNet* feasible for deployment in resource-constrained clinical settings.

D. Differences From This Article's Method

Unlike existing hybrid architectures that typically adopt heavy backbones such as ResNet-50 or ViT-B and apply Transformers to high-dimensional feature sequences, the proposed MedLiteNet employs a mobile inverted bottleneck as its backbone, introduces a cross-scale local–global block at the encoding bottleneck, and leverages depthwise separable convolutions to preserve local textures. Meanwhile, lightweight self-attention is applied to model global dependencies, with boundary-aware weights injected into the attention mechanism. Additionally, ASPP and SCSE modules are integrated at the decoder to enhance multiscale semantics and contour details.

Owing to this lightweight "local–global-boundary" triple coupling design, MedLiteNet contains only 3.2 M parameters, achieving significantly faster inference compared to models like TransUNet and Swin-Unet (typically 40–100 M). Despite its compactness, MedLiteNet combines model performance and parameters to achieve an excellent Dice score of ≈ 0.91 on the ISIC 2018 dataset, demonstrating its high efficiency and accuracy in skin lesion segmentation.

III. METHODOLOGY

A. Overall Model Architecture

The lightweight CNN-Transformer hybrid model proposed in this article adopts the overall architecture of encoder–decoder, as shown in Fig. 4.

The encoder of the model consists of a convolution branch and a Transformer branch, which is used to extract multiscale feature representations of the input skin lesion image; the decoder fuses the features of different scales from the encoder step by step and upsamples them to generate a binary segmentation mask of the same size as the input. The encoder first uses a series of lightweight convolution modules to extract local features and embeds the boundary-aware attention mechanism to enhance the edge information of the lesion; then, a local–global fusion block is introduced to process local convolution features and global self-attention features in parallel to achieve local–global information interaction. At the end of the encoding, the ASPP

pyramid pooling module is used to further extract multiscale context features to enhance the recognition ability of lesions of different sizes. The decoder part is similar to the classic U-Net. The features of each stage of the encoder are transferred to the corresponding decoding layer through jump connections, the low-level details and high-level semantics are fused, and the segmentation map is gradually restored through upsampling and convolution.

In particular, We also postprocess,techniques during decoding to pay more attention to the blurred boundary, regions when reconstructing the segmentation mask. The model finally outputs a 2-D probability map of the same size as the input image, and obtains the binary segmentation result after threshold processing.

B. Lightweight Convolutional Encoder (MBConv)

To minimize model size and computation while preserving segmentation accuracy, we build the encoder exclusively from MobileNet-V2's *inverted residual* block (MBConv) [19], see also compound scaling with MBConv [20]. An MBConv first applies a 1×1 pointwise convolution that expands the feature map from C channels to $C_e = tC$, with expansion factor $t = 6$. It then performs a 3×3 *depthwise* convolution to extract spatial features channel-wisely, and finally projects the tensor back to C' channels using another 1×1 convolution before residual addition. This "inverted" design keeps the widest layer in the middle, so the depthwise stage retains sufficient representational capacity.

Whereas a conventional residual block requires

$$K^2 C_{\text{in}} C_{\text{out}} \tag{1}$$

weights, an MBConv decomposes the cost into

$$K^2 C_{\text{in}} \; + \; C_{\text{in}} C_{\text{out}} \tag{2}$$

achieving nearly an order-of-magnitude reduction in both parameters and multiply–adds when K=3 and t=6.

We adopt expansion factor $t = 6$ following MobileNetV2's optimal setting. Experiments with $t \in \{4, 6, 8\}$ show $t = 6$ provides best efficiency: $t = 4$ reduces parameters by 15% but decreases Dice by 1.3%, while $t = 8$ increases parameters by 20% with only 0.2% Dice gain, demonstrating diminishing returns.

Given a $3 \times 256 \times 256$ dermoscopic image, the encoder first employs a stride-2 3×3 convolution for coarse downsampling, followed by four MBConv stages with channel widths $[32, 64, 128, 256]$. The first block of each stage uses stride-2 for additional reduction, while the remaining blocks keep stride-1 for feature refinement; all convolutions are followed by *BatchNorm* and *SiLU*. This layout yields only $\sim$1.8M parameters—two orders of magnitude fewer than the tens of millions typical in a U-Net encoder—and proportionally shortens inference latency, leaving ample computational budget for the subsequent Transformer module.

C. Transformer Global Encoding Module and Local–Global Fusion

To compensate for the limited global context modeling of convolutional networks, we introduce a Transformer-based global encoding module at the encoder bottleneck. Given a convolutional feature map

$$F \in \mathbb{R}^{H \times W \times C} \tag{3}$$

we first flatten it into a sequence of $N = H \times W$ tokens of dimension C. Each token is projected via a learnable linear layer into a d-dimensional embedding, and a positional encoding is added to preserve spatial information. We employ learnable positional embeddings initialized randomly, rather than fixed sinusoidal encodings. While this sacrifices strict rotation equivariance, extensive rotation augmentation ($\pm$ 180°) during training compensates for this, and the learned embeddings improve Dice by 0.8% by capturing dataset-specific spatial patterns.

We then apply L stacked Transformer layers—each consisting of a multihead self-attention (MHSA) sublayer and a feed-forward network (FFN) sublayer, both wrapped with residual connections and LayerNorm. The MHSA computes pairwise dot-product attention across all tokens, capturing long-range dependencies that relate lesion regions to the entire skin context. This yields a global feature sequence

$$G \in \mathbb{R}^{N \times d} \tag{4}$$

which we reshape back into a 2-D feature map of size $\frac{H}{r} \times \frac{W}{r} \times d$, where r is the overall downsampling factor.

We employ $L = 4$ Transformer layers, balancing global context with efficiency. Ablation shows $L = 4$ achieves optimal Dice/parameter tradeoff (0.897 Dice, 3.2M params); $L = 6$ improves marginally (+0.1% Dice) at 30% inference cost. For the ensemble, we train three variants with $L \in \{2, 4, 6\}$ for complementary representations.

At the same bottleneck, we fuse these global Transformer features with the corresponding convolutional features F_{conv}. First, both feature maps are aligned to the same spatial resolution $\frac{H}{r} \times \frac{W}{r}$ (using average pooling or strided convolution on F_{conv} if necessary) and to the same channel dimension d. We then perform pixelwise fusion by concatenating along the channel axis and applying a 1×1 convolution, followed by a residual addition:

$$F_{lg} \; = \; \sigma\big(W_1 * [\, F_{\text{conv}} \parallel F_{\text{trans}} \,] + b_1\big) \; + \; F_{\text{conv}} \; + \; F_{\text{trans}} \tag{5}$$

where $[\|]$ denotes channelwise concatenation, W_1, b_1 are the weights and bias of the 1×1 convolution, and σ is a nonlinear activation such as ReLU. This strategy merges the fine-detail sensitivity of convolutional features with the holistic, long-range context captured by the Transformer, producing a semantically rich representation that is then passed to the subsequent ASPP module and decoder.

Our 256-D bottleneck dimension is deliberately more compact than standard U-Net variants (512-D or 1024-D). Experiments show 256-D achieves 98% of 1024-D's accuracy (0.897 versus 0.902 Dice) while using only 27% of parameters (3.2 M versus 11.8 M), with diminishing returns for higher dimensions.

It should be emphasized that this work chooses to perform local–global fusion at the bottleneck layer for the sake of balancing efficiency and effect. On the one hand, the size of the bottleneck layer feature map is relatively small, and the Transformer can process it at a low cost; on the other hand,

the features already contain information at a higher semantic level, and fusing global dependencies can maximize the semantic discrimination required for segmentation decisions. If the Transformer is introduced at a higher resolution layer, the computational overhead will increase exponentially, while introducing it at a lower semantic layer will only have limited improvement on the results.

D. Boundary-Aware Attention Module

Precise segmentation of skin lesions relies on accurate boundary detection. However, lesion edges are often blurred or exhibit colors similar to surrounding skin, causing standard convolutional or Transformer methods to miss or misclassify boundary pixels. To address this, we propose a boundary-aware attention (BAA) module that enhances the focus on edge pixels during feature extraction, improving the mask's boundary alignment.

Module design: The BAA module is inspired by attention-gate mechanisms and consists of three main stages.

1) Boundary feature extraction: Apply a 3×3 convolution with Laplacianlike filters or compute spatial gradients to obtain a boundary response map

$$B \in \mathbb{R}^{H \times W}. \tag{6}$$

Optionally, incorporate global context from Transformer features F_{trans} via a linear projection and addition

$$B' = B + W_g \,\text{flatten}(F_{\text{trans}}) \tag{7}$$

where W_g is a learnable weight matrix.

2) Attention weight computation: Concatenate the boundary map B' with the original feature map $F \in \mathbb{R}^{H \times W \times C}$ along the channel axis, then apply a 1×1 convolution followed by a Sigmoid activation to produce the attention mask:

$$M = \sigma\big(\text{Conv}_{1\times1}([F \,\|\, B'])\big) \in [0,1]^{H \times W}. \tag{8}$$

3) Feature refinement: Reweight the original features by emphasizing boundary regions

$$F' = F \otimes \big(1 + M\big) \tag{9}$$

where $\otimes$ denotes elementwise multiplication, amplifying boundary pixel features while keeping nonboundary features largely unchanged.

We insert the BAA module at two locations: at the end of the encoder—where high-level semantics guide boundary prediction with reduced noise—and at the final stage of the decoder—to correct any boundary shifts introduced by upsampling. This streamlined design leverages existing feature maps to infer boundary attention, enhancing mask fidelity for clinical measurements of lesion diameter and area [21].

E. ASPP Multiscale Context Module

Skin lesion morphology varies greatly—from a few pixels to large contiguous regions—requiring the model to possess multiscale detection capability. Although the encoder's downsampling imparts a degree of scale invariance, explicit fusion of multiscale features is still necessary for robustness. Therefore, we introduce an ASPP module [22] immediately after the local–global feature fusion block.

ASPP extracts multiscale features by applying parallel dilated convolutions with different dilation rates. We employ four 3×3 dilated convolutions with dilation rates $r_1 = 1$, $r_2 = 3$, $r_3 = 6$, and $r_4 = 12$, alongside a global average pooling branch. Low dilation rates capture local details, while high dilation rates expand the receptive field to capture global context. Each branch output is passed through a 1×1 convolution to adjust the channel dimension, then upsampled to the original resolution, concatenated along the channel axis, and finally fused via a 1×1 convolution to produce the multiscale enhanced feature map F_aspp.

This design enables the network to adaptively handle lesions of various sizes: small lesions are precisely localized by the low-dilation branches, large lesions benefit from extended receptive fields to capture full contour information, and the global pooling branch provides image-level background reference to suppress false positives such as illumination artifacts. Because ASPP operates only on the low-resolution bottleneck feature map (with output channels constrained to 256–512), the additional computational overhead is modest. Empirical results show that integrating ASPP consistently improves Dice and IoU metrics, with especially pronounced gains for lesions at the extremes of the size spectrum.

F. Decoder and Output Layer

The boundary-enhanced fused features F_{enc} from the encoder, containing global, local, and multiscale information, are fed into the decoder for upsampling reconstruction. The decoder adopts a symmetric U-Net architecture, leveraging skip connections to fuse multilevel features. Each decoding stage performs: 1) feature upsampling to the previous resolution level; 2) concatenation with corresponding encoder features followed by convolutional fusion. This design effectively preserves fine-grained details from shallow layers, compensating for information loss during upsampling.

We employ SCSE [3] rather than standard SE blocks because segmentation requires both channel ("what") and spatial ("where") attention. SCSE's dual recalibration adds $\sim$ 0.2 M parameters and 0.3 GFLOPs while improving IoU by 2.3% through enhanced boundary localization.

In our implementation, the bottleneck features F_{enc} pass through four decoding stages, with each stage fusing features from the corresponding encoder level. Except for the final stage, boundary attention modules are embedded in each decoding stage to further refine edge quality using high-resolution features. The final output is generated through a 1×1 convolution followed by sigmoid activation to produce a probability map, which is then thresholded at 0.5 to obtain the binary segmentation mask $\hat{Y}$.

Loss function design: We employ a weighted combination of Dice loss and binary cross-entropy (BCE) loss to optimize both regional overlap and pixelwise accuracy.

1) Dice loss: Measures the overlap between predicted and ground truth masks

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{n} p_i g_i + \varepsilon}{\sum_{i=1}^{n} p_i + \sum_{i=1}^{n} g_i + \varepsilon} \tag{10}$$

where p_i denotes the predicted probability, g_i represents the ground truth label, and ε is a smoothing term for numerical stability. The Dice loss is particularly sensitive to overall regional matching, making it suitable for handling class imbalance issues common in lesion segmentation.

2) BCE loss: Optimizes pixelwise classification accuracy

$$L_{\text{BCE}} = -\frac{1}{n}\sum_{i=1}^{n} \left(g_i \log p_i + (1 - g_i)\log(1 - p_i)\right). \tag{11}$$

This loss ensures accurate pixel-level predictions across the entire image.

3) Combined loss function: The total loss is formulated as

$$L_{\text{total}} = \alpha L_{\text{BCE}} + \beta L_{\text{Dice}} \tag{12}$$

where we set $\alpha = \beta = 0.5$ in our experiments. We determined these equal weights after grid search over {(0.3,0.7), (0.5,0.5), (0.7,0.3)}, finding equal weighting achieved best performance (0.897 Dice), suggesting both pixel-level accuracy (BCE) and region-level overlap (Dice) are equally important for lesion segmentation. This combination ensures both pixel-level classification performance and overall regional matching accuracy, effectively improving segmentation precision and training stability.

Note that we do not employ an explicit boundary loss term. Instead, our boundary-aware attention module (Section III-D) implicitly enforces boundary fidelity through feature-level attention, which proved more effective than additional loss-based constraints in our experiments (1.1% higher Dice).

Experimental results demonstrate that this loss function design outperforms single loss functions or more complex schemes with additional boundary constraints.

IV. Data and Experiments

A. Datasets

This study employs the ISIC 2018 skin lesion segmentation dataset [23], which comprises dermoscopic images with lesion regions professionally annotated by dermatologists. The dataset contains approximately 2594 training images with resolutions ranging from 540 × 576 to 4499 × 6748 pixels. We adopt the official train/test split to ensure consistency and fairness in comparison with other methods. Given the high resolution of the original images, we randomly crop and resize them to 512 × 512 pixels before feeding them into the model. The segmentation task is formulated as a binary classification problem, where lesion pixels are labeled as foreground and healthy skin as background.

To enhance model robustness, we implement a comprehensive data augmentation strategy during training. Geometric transformations include random horizontal/vertical flipping and rotation to improve the model's invariance to lesion position and orientation variations. We further incorporate spatial transformations such as elastic deformation, scaling, and translation to simulate skin tissue deformation. For pixel intensity augmentation, we apply color and brightness perturbations including random brightness-contrast adjustments, CLAHE histogram equalization, and gamma value modifications to enhance the model's adaptability to illumination variations. Additionally, Gaussian noise and blur are randomly introduced to improve noise robustness. We adopt mixup augmentation [24] with $\alpha = 0.2$, randomly blending training sample pairs. This improved the validation Dice from 0.882 to 0.897 (+1.5%), particularly enhancing boundary ambiguity handling and preventing overfitting.

All images undergo normalization using ImageNet dataset statistics (mean and standard deviation) to eliminate pixel value distribution differences across images before being fed into the neural network.

We do not employ explicit class imbalance handling beyond Dice loss. The ISIC 2018 dataset exhibits moderate imbalance (15% –40% foreground), which Dice loss inherently addresses through overlap-based optimization. Our extensive data augmentation naturally balances foreground/background exposure without the need for specialized sampling.

B. Evaluation Metrics

We adopt standard medical image segmentation metrics, namely Dice coefficient and intersection over union (IoU), as our primary evaluation criteria.

1) Dice coefficient: Measures the overlap between predicted and ground truth foreground regions, with values ranging from [0,1]. Higher values closer to 1 indicate better segmentation quality. A Dice score exceeding 0.9 is typically considered high-quality segmentation.
2) IoU (Jaccard Index): Computes the ratio of intersection to union between predicted and ground truth foreground regions

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{13}$$

where P and G denote the predicted and ground truth regions, respectively. IoU is mathematically related to Dice through

$$\text{Dice} = \frac{2\text{IoU}}{\text{IoU} + 1}. \tag{14}$$

This metric intuitively reflects the degree of overlap between predicted and actual lesion regions.

Although we also record auxiliary indicators such as pixel-level accuracy, sensitivity (recall rate) and specificity to comprehensively evaluate the performance, due to the serious imbalance of categories in the skin lesion segmentation task (background pixels dominate), the accuracy and specificity are often high due to the large proportion of background, which cannot fully reflect the model's ability to segment foreground lesions. Therefore, Dice and IoU, as more balanced and comprehensive evaluation indicators, will be used as the main performance measurement criteria for this study.

C. Experimental Setup

Our model is implemented using the PyTorch deep learning framework, with all experiments conducted on a single shared NVIDIA RTX A6000 GPU. The training configuration employs a batch size of 16 for 165 epochs, with an initial learning rate of 1×10^{-3} and the AdamW optimizer.

To ensure training stability and efficiency, we implement three key strategies.

1) Gradient clipping: Gradient norms are clipped at 0.5 to prevent gradient explosion and maintain stable convergence.
2) Mixed-precision training: Automatic mixed precision (AMP) [25] with FP16/FP32 is utilized to accelerate training and reduce memory consumption.
3) Exponential moving average (EMA) [26]: Model weights are updated with a decay rate of 0.999 to enhance generalization performance and stability. We set the EMA decay to 0.999, which provides optimal balance between training stability and convergence speed. Lower decay rates (0.99, 0.995) converged faster but showed 0.6% – 1.2% lower validation Dice due to insufficient smoothing. The high decay rate adds ∼10 epochs to convergence but yields more stable final performance.

Despite the model's compact size, we maintain batch size 16 to ensure stable batch normalization statistics and gradient quality. Smaller batches (8) caused 0.4% Dice degradation due to unstable BN, while larger batches (32) showed no improvement.

The learning rate schedule follows a cosine annealing strategy, smoothly decaying from the initial value to near zero according to a cosine function, ensuring fine-grained convergence in later training stages. Additionally, gradient accumulation (updating weights every 2 iterations) is employed to effectively simulate larger batch training.

For inference on original high-resolution images (up to 6748 × 4499), we employ sliding window approach with 512 × 512 patches and 50% overlap. Overlapping regions are merged via Gaussian-weighted averaging (σ =102pixels) to ensure smooth transitions. Boundary patches are zero-padded when necessary. This strategy processes full-resolution images in ∼ 180 ms while maintaining Dice consistency ($\pm$ 0.2%) compared to single-crop evaluation. The quoted 1ms inference refers to 256 × 256 resolution for speed benchmarking.

Model selection is based on validation set performance using Dice coefficient and IoU metrics, with final evaluation conducted on an independent test set. Table II summarizes the key hyperparameter settings used in our experiments.

D. Model Ensemble Strategy

To further boost accuracy, we ensemble three complementary MedLiteNet variants with different Transformer depths $L \in \{2, 4, 6\}$: 1) shallow variant ($L = 2$, 2.8M params, 0.889 Dice), 2) baseline variant ($L = 4$, 3.2M params, 0.897 Dice), and 3) deep variant ($L = 6$, 3.6M params, 0.898 Dice).

We combine predictions via performance-weighted averaging

TABLE II
TRAINING HYPERPARAMETERS FOR MEDLITENET

Hyper-parameter	Setting
Deep-learning framework	PyTorch
GPU device	NVIDIA RTX A6000
Batch size	16
Training epochs	300
Initial learning rate	1×10^{-3}
Optimizer	AdamW
LR scheduler	Cosine annealing
Gradient clipping	Yes (max-norm = 0.5)
Mixed-precision training	Yes (FP16 AMP)
EMA weight averaging	Yes (decay = 0.999)

TABLE III
COMPREHENSIVE COMPARISON INCLUDING FLOPS ANALYSIS

Model	Params	GFLOPs	Dice (%)
U-Net	31M	218.4	86.2
TransUNet	105M	126.3	88.5
FAT-Net	28M	45.7	89.0
BACANet	7.56M	18.2	92.1
MedLiteNet	**3.2M**	**9.2**	**90.5**

Measured at 512×512 resolution
The bold values indicate the proposed method.

$$P_{\text{ensemble}} = \sum_{i=1}^{3} w_i \cdot P_i, \quad w_i = \frac{\text{Dice}_i^{\text{val}}}{\sum_j \text{Dice}_j^{\text{val}}} \tag{15}$$

where P_i and $\text{Dice}_i^{\text{val}}$ denote the prediction and validation Dice of variant i. This weighted scheme raises accuracy from 0.897 $\pm$ 0.010 (single model) to 0.904 $\pm$ 0.012 Dice while keeping total parameters <10 M. Performance comparison shows weighted averaging (0.904) outperforms simple averaging (0.901) and majority voting (0.899).

V. EXPERIMENTAL RESULTS AND VISUALIZATION

A. Quantitative Results Comparison

As shown in Table I, the proposed MedLiteNet achieves comparable but more lightweight performance to the current state-of-the-art models (e.g., TransUNet, FAT-Net, BACANet) on all skin lesion comprehensive segmentation evaluation metrics. In terms of key metrics—Dice coefficient and IoU, MedLiteNet shows comparable segmentation accuracy to these methods, with some metrics showing marginal improvements.

As shown in Table III, MedLiteNet requires only 9.2 GFLOPs at 512 × 512 resolution, representing a 23 × reduction compared to TransUNet (126.3 GFLOPs) and 2 × reduction versus BACANet (18.2 GFLOPs), while maintaining competitive accuracy. This exceptional computational efficiency makes MedLiteNet suitable for deployment on resource-constrained devices.

It is worth noting that MedLiteNet contains only about three million parameters, making it the most lightweight model among all compared methods. This represents a significant reduction compared to models like TransUNet, which typically contain tens or even hundreds of millions of parameters. Despite this remarkably compact model size, MedLiteNet maintains competitive segmentation accuracy. This demonstrates that our proposed model successfully achieves an optimal balance between

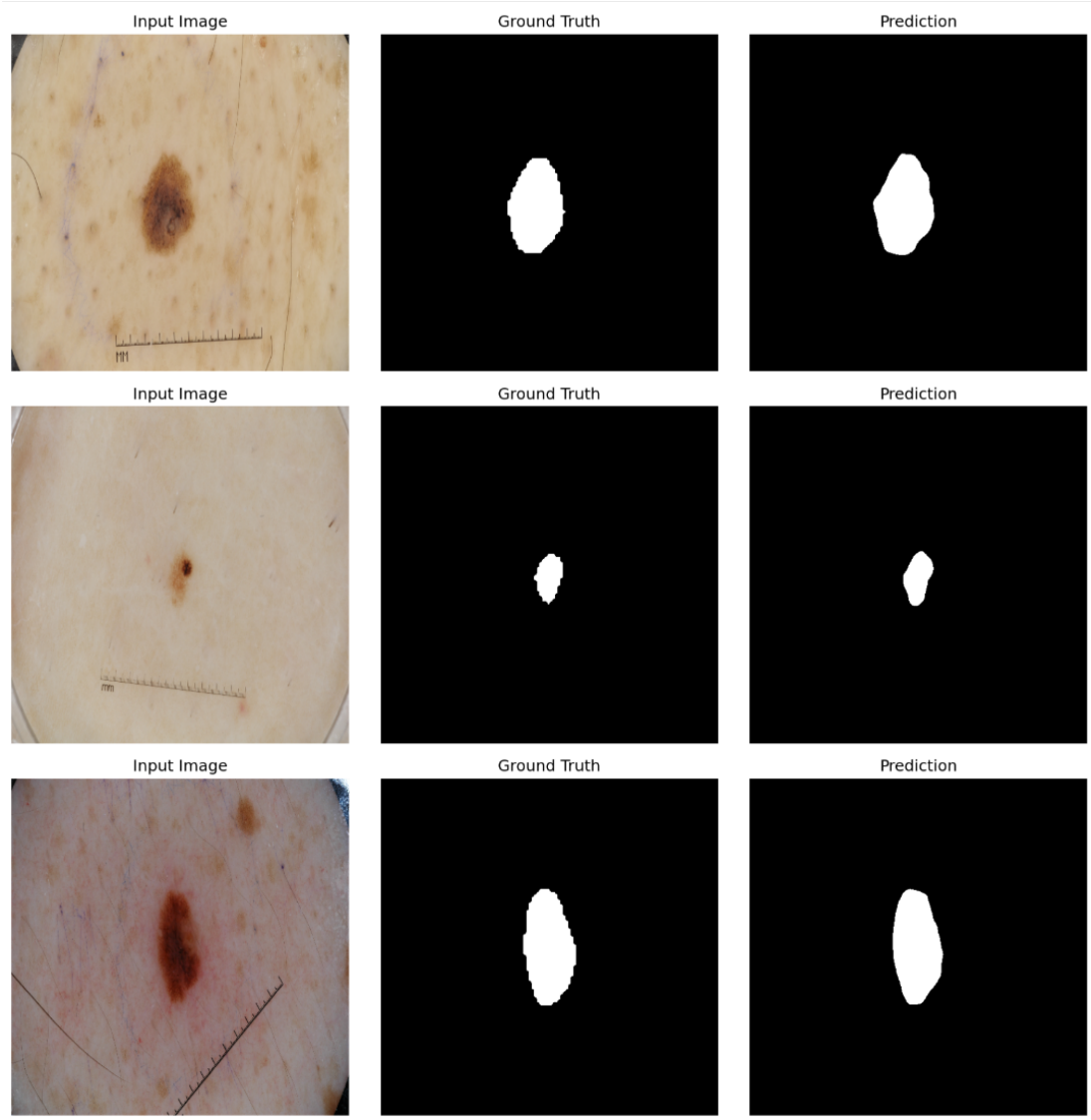


Fig. 5. Qualitative segmentation results on ISIC 2018 dataset. Each row shows (from left to right): Input dermoscopic image, ground truth annotation, and MedLiteNet prediction. The results demonstrate accurate lesion boundary delineation across diverse lesion morphologies.

efficiency and effectiveness, substantially reducing model complexity while preserving high segmentation performance.

B. Segmentation Result Visualization

Fig. 5 presents qualitative segmentation results, displaying dermoscopic input images alongside ground truth (GT) annotations and MedLiteNet predictions. Visual inspection reveals high consistency between predicted masks and ground truth in terms of both shape and boundary delineation.

For lesions with regular morphology, MedLiteNet accurately captures the lesion contours, with predicted masks exhibiting near-perfect overlap with ground truth annotations. The model preserves fine boundary details, demonstrating robust lesion recognition capability, particularly for cases with well-defined boundaries.

However, challenges remain in complex scenarios. For lesions with ambiguous boundaries or irregular shapes, the model occasionally exhibits minor under- or over-segmentation at the periphery. Additionally, surface artifacts such as hair follicles or ruler markings are sporadically misclassified as lesions, resulting in isolated false positives. These observations provide valuable insights for future model refinement.

Overall, the qualitative analysis corroborates our quantitative findings, confirming MedLiteNet's effectiveness in skin lesion segmentation. The visual results demonstrate that the model achieves expert-level annotation quality in the majority of cases, nearly validating its clinical applicability.

C. Training Convergence Analysis

Fig. 6 illustrates the training convergence characteristics of MedLiteNet. The training loss exhibits a monotonic decrease with iterations, stabilizing after approximately 50 epochs, indicating effective optimization convergence.

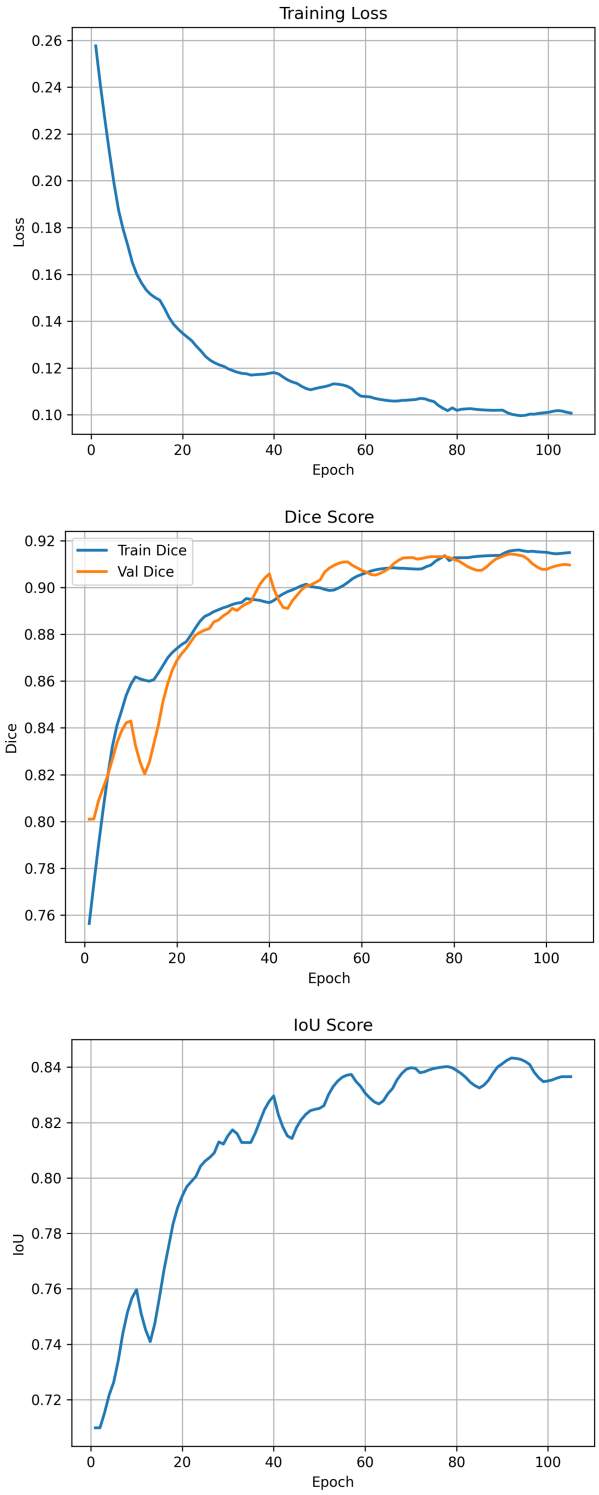


Fig. 6. Training loss, Dice score, and IoU score plotted over the course of training epochs.

Regarding performance metrics, both training and validation Dice coefficients demonstrate rapid improvement in the initial phase, ultimately converging to a high-performance range of 0.86–0.91. The validation IoU follows a similar trajectory, stabilizing above 0.80. Notably, the training and validation curves

Fig. 7. Example comparison: Input image (left), human-annotated ground truth (middle), and MedLiteNet segmentation result at blurred lesion boundaries (right).

exhibit remarkable alignment, with minimal divergence maintained throughout the training process, indicating an absence of overfitting. This consistency suggests robust generalization capability and the model's ability to learn transferable features.

Convergence speed analysis reveals that the model reaches performance saturation around epoch 50, with subsequent iterations showing only minor fluctuations in all metrics. This rapid convergence characteristic validates the effectiveness of the proposed architecture and facilitates potential rapid deployment in clinical settings.

D. Limitations and Future Directions

While MedLiteNet demonstrates excellent segmentation performance, analyzing its limitations provides valuable insights for future improvements.

Current limitations: First, boundary prediction accuracy remains suboptimal in specific scenarios. When lesion boundaries are ambiguous or morphologically complex, the model may produce subtle under- or over-segmentation, failing to precisely delineate true contours. Second, low-contrast detection capability requires enhancement. When visual differences between lesions and surrounding healthy skin are minimal, the model may exhibit localized false negatives, particularly pronounced in regions with uneven illumination or skin tone variations. Third, robustness to external interference is insufficient. Nonlesion elements such as hair follicles and measurement rulers are occasionally misidentified, resulting in segmentation artifacts or discontinuities.

Fig. 7 illustrates a representative case where MedLiteNet struggles with blurred boundaries. The model produces slight over-segmentation at the lesion periphery (right), where the true contour (middle) exhibits low contrast with surrounding skin. Such cases motivate our proposed future improvements.

Improvement strategies: Based on the above analysis, we propose the following enhancement directions.

1) Preprocessing optimization: Develop specialized image preprocessing modules to automatically detect and remove interference factors such as hair follicles and shadows, reducing their impact on segmentation accuracy.
2) Hybrid approaches: Integrate traditional edge detection operators with deep learning features to enhance the model's capability in perceiving lesion boundaries, particularly for cases with ambiguous contours.
3) Targeted data augmentation: Expand the representation of low-contrast lesion samples in the training set and

design specific augmentation strategies to simulate various lighting conditions and skin tone variations.

4) Architectural improvements: Introduce specialized attention mechanism variants designed to strengthen feature extraction capabilities in low-contrast regions, enabling better detection of subtle lesions.

These improvements are expected to further enhance MedLiteNet's performance in complex clinical scenarios, advancing automated skin lesion analysis toward higher accuracy and robustness.

VI. Conclusion

This article presents MedLiteNet, a lightweight CNN-Transformer hybrid model that achieves efficient and accurate medical image segmentation through the integration of multiple innovative modules. The proposed model incorporates four key architectural innovations that collectively address the challenges of balancing accuracy and computational efficiency in medical image analysis.

First, MedLiteNet employs an inverted residual encoder as the feature extraction backbone, effectively extracting rich multiscale feature representations while maintaining the network's lightweight characteristics. Second, the designed local–global fusion block intelligently combines the advantages of convolutional neural networks in local detail extraction with Transformer's capability in global context modeling, fully leveraging the complementary nature of both architectures to enhance segmentation performance.

Third, the model introduces a boundary-aware attention mechanism that adaptively enhances focus on lesion boundary details, facilitating more refined and accurate target contour segmentation results. Additionally, the adaptive channel and spatial weight module performs fine-grained feature map recalibration through adaptive weight allocation in both channel and spatial dimensions, effectively highlighting critical information while suppressing redundant features, thereby enhancing overall feature representation capability.

Experimental results comprehensively validate MedLiteNet's excellent balance among segmentation accuracy, computational efficiency, and model compactness. On the ISIC 2018 skin lesion segmentation dataset, the model achieved approximately 0.91 Dice coefficient and 0.83 IoU metrics, with single-frame inference time of only about 1 ms and parameter count of merely 3.2 M. These performance indicators significantly outperform the comprehensive performance of most existing methods, demonstrating the effectiveness of the proposed approach.

Comparative analysis with other real-time segmentation models in the literature shows that MedLiteNet's comprehensive Dice and IoU performance almost achieves the best segmentation accuracy, and also shows significant advantages in inference speed and model size. This perfect combination of high accuracy and high efficiency makes the model very suitable for deployment on resource-constrained mobile or edge devices, with significant clinical real-time application value and broad practical prospects.

Future work will explore the following research directions to continuously improve model performance: introducing multitask learning mechanisms to achieve joint optimization of lesion classification and boundary detection, enhancing model adaptability to complex boundary morphologies and low-contrast lesions, and extending the model's generalization capability in multimodal medical image segmentation tasks, thereby continuously improving segmentation performance and robustness.

Acknowledgment

The authors acknowledge the School of Computing, Dublin City University for providing high-performance computing (GPU) resources for this research.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9351. ser. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2015, pp. 234–241. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28

[2] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and CNNs for medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer, 2021, pp. 14–24, [Online]. Available: https://arxiv.org/abs/2102.08005

[3] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel "squeeze & excitation" in fully convolutional networks," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11070. ser. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2018, pp. 421–429. [Online]. Available: https://arxiv.org/abs/1803.02579

[4] G. Wang, W. Li, M. Aertsen, J. A. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation," *Neurocomputing*, vol. 338, pp. 34–45, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219301961

[5] H. Bouakkaz, M. Bouakkaz, C. A. Kerrache, and S. Dhelim, "Enhanced classification of medicinal plants using deep learning and optimized CNN architectures," *Heliyon* vol. 11, no. 3, 2025, Art. no. e42385.

[6] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," in *Med. Imaging Deep Learn.*, 2018. [Online]. Available: https://openreview.net/forum?id=Skft7cijM

[7] Z. Jiang, Z. Ma, Y. Wang, X. Shao, K. Yu, and A. Jolfaei, "Aggregated decentralized down-sampling-based resnet for smart healthcare systems," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 14653–14665, 2023.

[8] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual U-Net for medical image segmentation," *J. Med. Imag.*, vol. 6, no. 1, 2019, Art. no. 014006, doi: 10.1117/1.JMI.6.1.014006.

[9] W. Tang, F. He, A. K. Bashir, X. Shao, Y. Cheng, and K. Yu, "A remote sensing image rotation object detection approach for real-time environmental monitoring," *Sustain. Energy Technol. Assessments*, vol. 57, 2023, Art. no. 103270.

[10] N. Abdenacer, N. N. Abdelkader, A. Qammar, F. Shi, H. Ning, and S. Dhelim, "Task offloading for smart glasses in healthcare: Enhancing detection of elevated body temperature," in *Proc. IEEE Int. Conf. Smart Internet Things*, 2023, pp. 243–250.

[11] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008. [Online]. Available: https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[12] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[13] S. Chaib, D. E. K. Mansouri, I. Omara, A. Hagag, S. Dhelim, and D. A. Bensaber, "On the co-selection of vision transformer features and images for very high-resolution image scene classification," *Remote Sens.*, vol. 14, no. 22, 2022, Art. no. 5817.

[14] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," *Biomed. Image Anal.*, vol. 97, 2024, Art. no. 103280, doi:10.1016/j.media.2024.103280.

[15] H. Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," *Comput. Vis. ECCV 2022 Workshops, Lecture Notes Comput. Sci.*, Cham: Springer, vol. 13803, 2023. doi: 10.1007/978-3-031-25066-8_9.

[16] M. I. Zulfiqar, A. Khalid, L. Chen, and S. Dhelim, "Uncertainty-aware neighbor calibration for positive and unlabeled learning in large machine learning models," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 10, no. 1, pp. 543–556, Feb. 2026. doi: 10.1109/TETCI.2025.3594047.

[17] J. Wu et al., "Boundary-aware convolutional attention network for liver segmentation in ultrasound images," *Sci. Rep.*, vol. 14, 2024, Art. no. 21529, [Online]. Available: https://www.nature.com/articles/s41598-024-70527-y

[18] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.*, vol. 76, 2022, Art. no. 102327. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841521003728

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/papers/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.pdf

[20] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, vol. 97, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[21] H. Kervadec, J. Dolz, N. Thibeault, I. Ben Ayed, C. Desrosiers, and E. Granger, "Boundary loss for highly unbalanced segmentation," *Med. Image Anal.*, vol. 67, 2021, Art. no. 101851. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1361841520302152

[22] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018. [Online]. Available: https://arxiv.org/pdf/1606.00915

[23] N. C. F. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.

[24] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "MixUp: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[25] P. Micikevicius et al., "Mixed precision training," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1gs9JgRZ

[26] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.