Enhancing Early Skin Cancer Detection: A Deep Learning Approach with Multi-Scale Feature Refinement and Fusion

Siyuan Wu , Pengfei Zhao , Huafu Xu , Ziming Wang *

Posted Date: 2 March 2026

doi: 10.20944/preprints202603.0154.v1

Keywords: skin cancer; skin lesion segmentation; multi-scale encoder feature;adaptive offset prediction

 doi:10.20944/preprints202603.0154.v1

Article

Enhancing Early Skin Cancer Detection: A Deep Learning Approach with Multi-Scale Feature Refinement and Fusion

Siyuan Wu [1], Pengfei Zhao [1], Huafu Xu [2,3] and Ziming Wang [4,*]

[1] Guilin Institute of Information Technology, Guilin, Guangxi 541004 China
[2] Graduate School of Education, Peking University, Beijing, 100871, China
[3] Guangxi Zhuang Autonomous Region Bureau of Big Data Development, Nanning, 530201, China
[4] Guilin University of Electronic Technology, School of Computer Science and Information Security, Guilin, Guangxi 541004 China
* Correspondence: worthyman@guet.edu.cn

Abstract

The global incidence of skin cancer is rising, making it an increasingly critical public health issue. Malignant skin tumors such as melanoma originate from pathological alterations of skin cells, and their accurate early-stage segmentation is crucial for quantitative analysis, early diagnosis, and successful treatment. However, achieving precise and efficient segmentation remains a major challenge, as existing methods often struggle to balance computational efficiency with the ability to capture complex lesion characteristics. To address this challenge, we propose a novel deep learning framework that integrates the PVT v2 backbone with two key modules: Spatial-Aware Feature Enhancement (SAFE) and Multiscale Dual Cross-attention Fusion (MDCF). The SAFE module refines multi-scale encoder features through a dual-branch architecture that bridges the feature discrepancy across network depths by combining fine-grained shallow-layer details with deep semantic information via adaptive offset prediction. The MDCF module establishes bidirectional cross-attention between decoder and encoder features, followed by multi-scale deformable convolutions that capture lesion boundaries and small fragments at heterogeneous receptive fields, thereby enriching semantic details while suppressing background responses. The proposed model was evaluated on two public benchmark datasets (ISIC 2016 and ISIC 2018), achieving Intersection over Union (IoU) scores of 87.33% and 83.67%, respectively, demonstrating superior performance compared to current state-of-the-art methods. These results indicate that our framework significantly enhances skin lesion image analysis and offers a promising tool for improving early detection of skin cancer.

Keywords: skin cancer; skin lesion segmentation; multi-scale encoder feature;adaptive offset prediction

1. Introduction

In recent years, environmental pollution and increased ultraviolet radiation have contributed to a significant rise in the incidence of skin cancers such as melanoma, drawing considerable attention from the global medical community [1,2]. Early diagnosis is crucial for improving treatment outcomes and patient survival rates [2,3]. Traditional diagnostic methods, which primarily rely on clinical observation and tissue biopsy, are subjective, invasive, and unsuitable for large-scale screening [4,5]. Medical image segmentation technology offers a non-invasive, high-precision alternative for skin cancer analysis, providing clinicians with richer and more accurate diagnostic information [6,7] and holding promise for breakthroughs in early detection and treatment.

Automatic segmentation of skin lesions plays a vital role in computer-aided diagnosis systems for melanoma [6,7]. Data from 2022 reveal significant geographical disparities in skin cancer incidence, with Australia reporting 37 cases per 100,000 people, China reporting only 0.37 cases, and a global average of 3.2 cases [1,2]. These variations highlight the need for adaptable diagnostic tools that can

 doi:10.20944/preprints202603.0154.v1

perform effectively across diverse populations and clinical settings. Public datasets such as ISIC have become critical resources for developing and validating such technologies, enabling researchers to train robust models that generalize across different demographic and geographic contexts [14,15].

Dermoscopy, a non-invasive magnification imaging technique, has emerged as a key tool in melanoma diagnosis [4,5]. Studies indicate that early diagnosis can achieve 10-year survival rates of 84% to 98%, whereas late-stage diagnosis results in only 10% to 15% survival [3]. However, manual interpretation of dermoscopic images is time-consuming and labor-intensive, underscoring the necessity of automatic segmentation technology to enhance diagnostic accuracy and efficiency [6,7]. Automatic skin lesion segmentation aims to perform pixel-level annotation of lesion areas in dermoscopic images. However, this task faces substantial challenges, including variations in skin color and lesion size, texture diversity, and artifacts such as shadows and body hair [6,9]. Early segmentation techniques, such as thresholding and color space analysis, relied on predefined feature extraction and often failed to handle the variability in lesion appearance and interference from external artifacts [8,9].

Against the backdrop of the continuously rising global incidence of skin cancer, achieving precise early segmentation of malignant skin tumors such as melanoma has become critical for enhancing quantitative analysis accuracy, assisting early clinical diagnosis, and optimizing treatment strategies. However, existing segmentation methods face the fundamental challenge of balancing computational efficiency with feature capture capability, limiting their effectiveness in practical clinical applications.Traditional methods based on threshold segmentation, edge detection, or color-texture feature extraction can achieve preliminary lesion identification [8]. However, they are highly dependent on image quality and struggle to handle interfering factors such as lesion morphological diversity, blurred boundaries, and artifacts (e.g., shadows, hair follicles) [6,9]. Although deep learning models such as U-Net have significantly improved segmentation performance, they still exhibit limitations in multi-scale feature fusion, long-range contextual modeling, and detail recovery [10,11]. These limitations are particularly evident in cases involving discontinuous boundaries or small lesions, where under-segmentation and mis-segmentation frequently occur [11].

To address these shortcomings, this paper proposes a novel deep learning segmentation framework that integrates a Pyramid Vision Transformer v2 (PVT v2) backbone with two collaborative modules—Spatial-Aware Feature Enhancement (SAFE) and Multi-scale Dual Cross-attention Fusion (MDCF) to enhance feature representation and improve decoding reconstruction [12,13]. As illustrated in Figure 1, we benchmarked our proposed method against current state-of-the-art approaches on the ISIC 2016 and ISIC 2018 datasets. The results demonstrate notable enhancements in both the IOU and DICE metrics. The core advantages of this method are reflected in three key aspects:

First, the PVT v2 encoder leverages its pyramidal hierarchical structure to effectively extract lesion features at multiple scales [13]. This multiscale representation enables the model to adapt to lesions of varying sizes and shapes, from small isolated spots to large irregular regions, thereby improving robustness across diverse clinical scenarios.

Second, the SAFE module refines encoder features at each resolution level through a dual-branch architecture that jointly leverages shallow and deep representations. The base-feature branch preserves fine-grained spatial details from shallow layers, while the offset-compensation branch employs dilated convolutions to predict adaptive offsets that align shallow features with deep semantic information. This mechanism effectively bridges the feature discrepancy across network depths, improving the discriminability and completeness of multi-level feature representation.

Third, the MDCF module establishes bidirectional cross-attention between decoder and SAFE-enhanced encoder features to enable dynamic semantic information transfer via skip connections. It employs a four-branch multi-scale deformable convolution unit to capture lesion boundaries and small fragments at heterogeneous receptive fields. This design significantly improves boundary segmentation accuracy and fine structure recovery while suppressing background responses. In summary, the main contributions of this paper are as follows:

 doi:10.20944/preprints202603.0154.v1

- **Novel Architecture with PVT v2 Backbone:** We propose a new deep learning framework for automatic skin lesion segmentation that incorporates the Pyramid Vision Transformer v2 (PVT v2) as the core encoder. This design effectively captures multi-scale contextual features from dermoscopic images, enhancing the model's ability to recognize lesions of varying sizes and complexities [13].
- **Dual-Module Cooperative Mechanism:** The framework introduces a synergistic dual-module mechanism within the decoder pathway. The first module is dedicated to the sophisticated integration of low-level spatial details and high-level semantic features, ensuring a more discriminative and complete feature representation. The second module employs an enhanced skip-connection strategy for dynamic feature fusion between the encoder and decoder, significantly improving the precision of boundary delineation and the recovery of fine-grained structures [10,12].
- **Comprehensive Evaluation on Benchmark Datasets:** We conduct extensive experiments on publicly available datasets, such as ISIC 2016 and SISC 2018. The results demonstrate that our proposed method achieves state-of-the-art performance, particularly in handling boundary ambiguity and small lesions, outperforming existing leading approaches in terms of standard segmentation metrics like Dice Coefficient and IOU. This validates the practical effectiveness and robustness of our model in a clinical simulation setting.

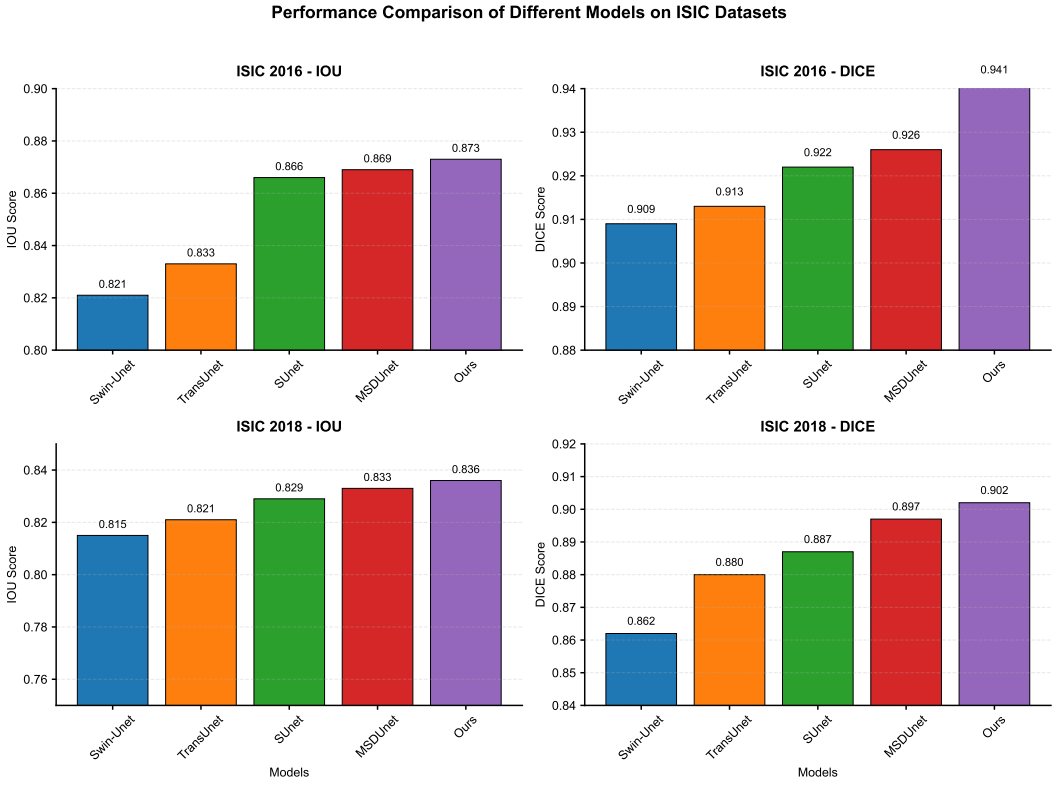

Figure 1. Performance Comparision of Different Models on ISIC Datasets.

2. Related Work

2.1. CNNs for Skin Lesion Segmentation

Convolutional Neural Networks (CNNs) have served as the cornerstone and dominant paradigm in automated skin lesion segmentation, revolutionizing the field through their ability to automatically learn hierarchical features from raw pixel data. The evolution of CNN-based approaches can be largely traced through the adoption and refinement of encoder-decoder architectures [9,10,29,30].

Among these, the U-Net architecture [10] has arguably been the most influential model in medical image segmentation, including skin lesion analysis. Its symmetric encoder-decoder structure, combined with skip connections, effectively addresses the challenge of integrating high-resolution

 doi:10.20944/preprints202603.0154.v1

spatial information from the encoder pathway with the high-level semantic information recovered during decoding. Skip connections are particularly critical for preserving fine-grained details along lesion boundaries, which are often irregular and poorly defined. Following U-Net, variants such as SegNet [16]—which utilizes pooling indices for upsampling in the decoder—and other encoder-decoder frameworks have been widely applied, cementing this architectural paradigm as the standard backbone for the task.

Building upon the encoder-decoder foundation, researchers have introduced various modifications to tackle challenges specific to dermoscopic images. While architectures like U-Net set the standard under full supervision, later work has sought to reduce dependence on costly pixel-level annotations. Weakly supervised methods, for instance, have attracted growing interest. A notable contribution by Anggasta Aji Azhari et al. [17] introduced a two-stage CNN that integrates Grad-CAM for weak localization to refine feature learning. This marks a conceptual shift: rather than using explainability tools only for post-hoc analysis, they are embedded directly into training to improve model confidence and interpretability—bridging high discriminative performance with clinical practicality. Beyond supervision strategies, a key challenge has been enhancing the ability of CNNs to capture long-range contextual information while preserving local detail. While recent efforts have incorporated self-attention mechanisms like Transformers, an emerging direction involves state space models. In this vein, BEFNet [20] employs a Mamba-based encoder to capture global context, while retaining a dedicated CNN branch to preserve local spatial precision—ensuring that critical boundary details are not lost. This reflects a design philosophy of complementary integration, where CNNs play a specialized role alongside global modeling approaches. Furthermore, the computational cost of deep models remains a barrier to clinical deployment. To address efficiency, Chun et al. [21] proposed UCM-Net, which combines MLPs with CNNs in a lightweight architecture requiring under 50KB parameters and 0.05 GLOPS, yet maintains competitive segmentation accuracy. This makes it a promising benchmark for resource-conscious applications in dermatology.

Another significant challenge in skin lesion segmentation is the severe class imbalance between small lesion regions and extensive background areas. To mitigate this issue, loss functions such as Dice Loss [18] and Focal Loss [19] have been widely adopted as alternatives to standard cross-entropy. These losses explicitly encourage the model to focus on underrepresented lesion pixels, thereby improving segmentation accuracy for the target structures.

2.2. Transformers for Skin Lesion Segmentation

With the rapid advancement of Vision Transformers (ViTs) [33], transformer-based architectures have emerged as powerful alternatives to traditional CNNs in medical image segmentation. Unlike convolutional networks that inherently focus on local spatial relationships, Transformers model long-range dependencies through self-attention, enabling more comprehensive global context understanding—an ability particularly beneficial for skin lesion segmentation, where lesions often present irregular textures and ambiguous boundaries. This advantage has been validated by the Medical Transformer [34], which introduces gated axial-attention to efficiently capture long-range dependencies and has demonstrated strong performance across multiple medical image segmentation tasks.

Hybrid architectures such as TransUNet [12] combine CNN encoders with transformer blocks to incorporate both local feature extraction and global dependency modeling, resulting in improved performance on medical imaging benchmarks. Similarly, hierarchical transformer models like Swin-UNet [22], built upon the shifted-window self-attention mechanism originally proposed in Swin Transformer [35], achieve efficient global modeling while maintaining linear computational complexity with respect to image size.

Pyramid-style transformers further push the performance for dense prediction tasks. The Pyramid Vision Transformer (PVT) and its improved variant PVT v2 [13] utilize spatial-reduction attention and multi-scale feature hierarchies, addressing computational bottlenecks while preserving global contextual cues—making them well-suited for high-resolution dermoscopic images. Meanwhile,

 doi:10.20944/preprints202603.0154.v1

segmentation frameworks such as SegFormer [23] adopt lightweight pyramid Transformers with MLP decoders, demonstrating robust generalization across diverse medical datasets.

Recent studies also incorporate cross-attention modules in the decoder to enhance dynamic feature fusion, improving boundary localization and facilitating better integration of multi-level semantic features. Overall, transformer-based approaches significantly outperform conventional CNNs in capturing long-range structure and resolving fuzzy lesion boundaries, and have thus become a key trend in skin lesion segmentation.

3. Materials and Methods

3.1. Architectural Overview

As illustrated in Figure 2, The decoder reconstructs a high-resolution segmentation map from these hierarchical features through a multi-stage refinement process. Starting from the deepest encoder output (Stage 4), feature maps are progressively upsampled and refined at each decoder level. At each resolution, the decoder features are enriched through two complementary mechanisms.

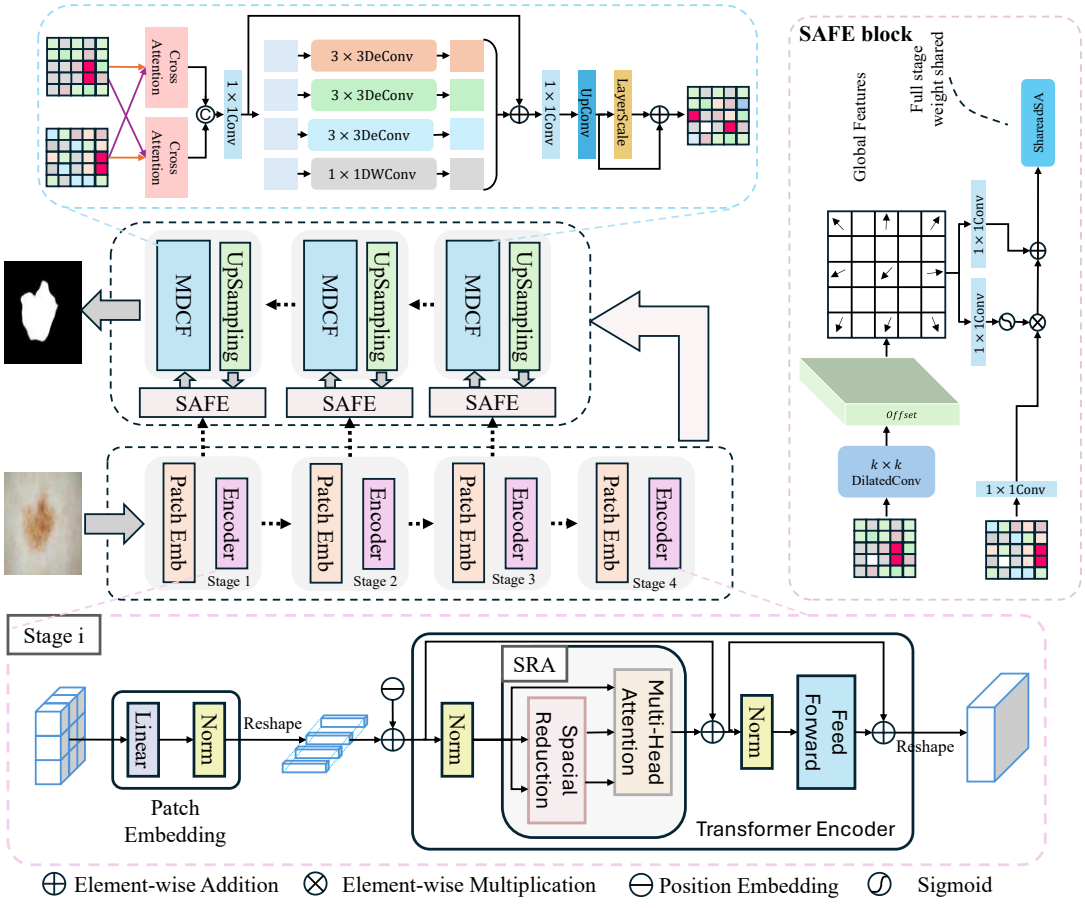

Figure 2. The proposed model architecture.

Each lateral connection from the encoder is processed by a SAFE module, which employs a dual-branch architecture to bridge the semantic gap between encoder and decoder features. The base-feature branch preserves fine-grained spatial details from shallow encoder layers, while the offset-compensation branch uses dilated convolutions and global context to predict adaptive spatial offsets, aligning shallow features with deep semantic information before fusion. After receiving SAFE-enhanced encoder features, each decoder stage applies the MDCF module to establish bidirectional cross-attention between encoder and decoder representations. This is followed by a four-branch multi-scale deformable convolution unit that captures lesion boundaries and small fragments at

 doi:10.20944/preprints202603.0154.v1

heterogeneous receptive fields, enabling adaptive feature aggregation while suppressing background responses.

A series of convolutional refinement layers and a lightweight prediction head convert the final high-resolution features into a pixel-wise probability map. This design synergistically combines the Transformer's global context modeling with SAFE-guided multi-scale fusion and MDCF-based adaptive aggregation to preserve fine structures while maintaining robust semantic consistency.

3.2. Encoder Model

The proposed network adopts a hierarchical Transformer-based encoder to extract multi-scale representations from the input dermoscopic image. As illustrated in Figure 2, the encoder is organized into four stages. Each stage consists of a patch embedding layer followed by a stack of Transformer blocks, and produces a feature map with reduced spatial resolution but increased channel dimension. The resulting feature maps are later used both to drive the decoder and as inputs to the SAFE modules.

At stage i, the input feature map $X_{i-1} \in \mathbb{R}^{H_{i-1} \times W_{i-1} \times C_{i-1}}$ is first processed by a linear projection implemented as a convolutional patch embedding layer. This layer uses a kernel size and stride larger than 1 to merge neighboring pixels into non-overlapping (or slightly overlapping) patches and to down-sample the spatial resolution. A layer normalization is then applied, and the 2-D feature map is reshaped into a sequence of $N_i = H_i W_i$ tokens of dimension C_i. To preserve spatial layout information, we add a learnable 2-D positional embedding to the token sequence before feeding it to the Transformer blocks.

Each Transformer block follows a pre-norm architecture and is equipped with spatial-reduction attention (SRA) to capture long-range dependencies with manageable computational cost. Specifically, the input tokens are first normalized and then fed into an SRA module. In SRA, the query tokens are kept at full resolution, while the key and value tokens are obtained from a spatially down-sampled version of the feature map, enabling global receptive fields with reduced quadratic complexity. The output of the multi-head attention is added to the input via a residual connection. A second normalization layer followed by a position-wise feed-forward network (two linear layers with a non-linearity in between) is then applied, again with a residual connection. Finally, the token sequence is reshaped back to a 2-D feature map to serve as input to the next encoder stage. This process can be expressed by the following formula:

$$T_0^{(i)} = \text{Reshape}\big(\text{Norm}(\text{Linear}(X^{(i)}))\big) \tag{1}$$

$$T_{l+\frac{1}{2}}^{(i)} = T_0^{(i)} + \text{MHA}_{\text{SRA}}\Big(\text{Norm}\Big(T_0^{(i)}\Big)\Big) \tag{2}$$

$$T_{l+1}^{(i)} = T_{l+\frac{1}{2}}^{(i)} + \text{FFN}\Big(\text{Norm}\Big(T_{l+\frac{1}{2}}^{(i)}\Big)\Big) \tag{3}$$

$$Y^{(i)} = \text{Reshape}^{-1}\Big(T_{l+1}^{(i)}\Big) \tag{4}$$

where $X^{(i)}$ denotes the input feature map of the i-th stage, $T_0^{(i)}$ is the token sequence obtained by linearly projecting, normalizing and reshaping $X^{(i)}$, $T_{l+\frac{1}{2}}^{(i)}$ and $T_{l+1}^{(i)}$ are the intermediate features after spatial-reduction multi-head self-attention and the subsequent feed-forward network at the l-th layer, respectively, $Y^{(i)}$ is the output feature map of this stage, $Norm(\cdot)$ denotes layer normalization, $MHA_{\text{SRA}}(\cdot)$ denotes multi-head self-attention with spatial reduction, $FFN(\cdot)$ denotes a position-wise feed-forward network, $Reshape(\cdot)$ and $Reshape^{-1}(\cdot)$ convert between 2D feature maps and 1D token sequences.

The encoder outputs from all stages are forwarded to the decoder through SAFE blocks. For a given stage, the corresponding feature map is fed into a SAFE block that exploits both global and local cues. A $k \times k$ dilated convolution is first applied to generate an offset-enriched representation capturing a larger receptive field. In parallel, the global features of the same stage are processed by

 doi:10.20944/preprints202603.0154.v1

a weight-shared 1×1 convolution. The two branches are fused and passed through another 1×1 convolution followed by a sigmoid activation to produce a spatial attention map. This attention map is then multiplied element-wise with the original stage feature map, yielding an enhanced representation that emphasizes lesion-related regions while suppressing irrelevant background responses. These SAFE-refined features at multiple scales provide rich contextual information for the subsequent decoder and segmentation head.

3.3. SAFE Model

The SAFE module is designed to refine the encoder features at each resolution level by jointly leveraging shallow-layer and deep-layer representations, while explicitly suppressing background responses. As shown in Figure 2, one SAFE block is attached to the output of each encoder stage, and the resulting enhanced features are forwarded to the decoder.

Given an input feature map from stage (i), denoted as $\mathbf{F}_i \in \mathbb{R}^{H_i \times W_i \times C_i}$, SAFE first prepares the features and then computes a stage-specific attention map. To reduce computational cost and stabilize training, $\mathbf{F}_i$ is optionally passed through a 1×1 convolution and a normalization layer to align the channel dimension. The processed feature map is then fed into two parallel branches: a base-feature branch and an offset-compensation branch. The base-feature branch preserves the shallow representation at stage (i), while the offset-compensation branch predicts offsets to compensate for the discrepancy between the shallow features and the corresponding deep features.

In the base-feature branch, a 1×1 convolution is applied to $\mathbf{F}_i$ to obtain a refined shallow representation $\mathbf{F}_i^{\text{base}}$. This projection mainly adjusts the channel dimension and preserves the stage-(i) content, so that low-level cues and fine details are maintained in a compact form. A normalization layer and a non-linear activation are used to stabilize training and further refine $\mathbf{F}_i^{\text{base}}$, providing a clean base feature for subsequent fusion with deeper layers.

In the offset-compensation branch, SAFE estimates a bias field that aligns the shallow features with their deep counterparts by jointly capturing local context and global statistics. First, a $k \times k$ dilated convolution is applied on $\mathbf{F}_i$ to obtain a locally enriched representation $\mathbf{F}_i^{\text{loc}}$. The dilation rate is chosen according to the spatial scale of the current stage, so that shallow stages focus on fine details while deeper stages capture larger structures. This operation increases the effective receptive field without reducing resolution, thereby aggregating neighboring cues that are important for lesion boundaries and texture. A 1×1 convolution and activation function then project $\mathbf{g}_i$ into an offset vector that is broadcast back to the spatial size $H_i \times W_i$ to produce an offset map $\mathbf{F}_i^{\text{off}}$. This offset map is used to modulate the base features, providing a learned bias that compensates for the discrepancy between shallow and deep features.

The outputs of the two branches are then fused to generate a stage-aware representation. Specifically, the offset-compensation branch output $\mathbf{F}_i^{\text{off}}$ is first passed through a non-linear activation function and used to modulate the base features via element-wise multiplication, yielding an intermediate feature $\mathbf{F}_i^{\text{mod}} = \sigma(\mathbf{F}_i^{\text{off}}) \odot \mathbf{F}_i^{\text{base}}$. This modulated feature is then added back to the original offset feature in a residual manner, i.e., $\mathbf{F}_i^{\text{res}} = \mathbf{F}_i^{\text{off}} + \mathbf{F}_i^{\text{mod}}$, and fed into the ShareadSA module to produce the final stage-aware output $\hat{\mathbf{F}}_i = \text{ShareadSA}(\mathbf{F}_i^{\text{res}})$. This fusion strategy enables the learned offsets to gate the shallow base features while preserving offset information, leading to stage-specific refinement for lesion segmentation.

$$\mathbf{F}_i^{\text{off}} = \text{Conv}_{1\times1}(\text{DilatedConv}_{k\times k}(\mathbf{F}_i)) \tag{5}$$

$$\mathbf{F}_i^{\text{base}} = \text{Conv}_{1\times1}(\mathbf{F}_i) \tag{6}$$

$$\hat{\mathbf{F}}_i = \text{ShareadSA}((\sigma(\mathbf{F}_i^{\text{off}}) \odot \mathbf{F}_i^{\text{base}} + \mathbf{F}_i^{\text{base}})) \tag{7}$$

where $\mathbf{F}_i$ denotes the input feature map at stage i, $\mathbf{F}i^{\text{off}}$ and $\mathbf{F}i^{\text{base}}$ denote the features of the offset branch and the base branch, respectively; DilatedConv$k \times k(\cdot)$ and Conv$1 \times 1(\cdot)$ denote a $k \times k$ dilated convolution and a 1×1 convolution, respectively; $\sigma(\cdot)$ denotes the sigmoid activation function, $\odot$

 doi:10.20944/preprints202603.0154.v1

denotes element-wise multiplication, and ShareadSA$(\cdot)$ denotes the shared stage-aware attention module.

3.4. Decoder Model

The decoder adopts a cascaded architecture that alternates MDCF blocks with Upsampling blocks, progressively transforming coarse high-level representations into fine-grained, high-resolution predictions. Three MDCF modules are arranged in sequence, each followed by an Upsampling operation (except at the finest scale). At every stage, the MDCF block receives the current decoder feature and its scale-consistent SAFE-enhanced encoder feature, so that spatial resolution is increased while scale-specific encoder information is injected throughout the decoding process.

At the i-th decoding stage, we denote the input decoder feature as $\mathbf{X}_i$ and the corresponding SAFE-refined encoder feature as $\tilde{\mathbf{F}}_i$. The MDCF block first establishes a bidirectional interaction between these two feature streams via a cross-attention mechanism, as illustrated in Figure 2. Specifically, we compute two directional cross-attention responses,

$$\mathbf{Z}_{d\leftarrow e} = \mathrm{CA}(\mathbf{X}_i, \tilde{\mathbf{F}}_i), \quad \mathbf{Z}_{e\leftarrow d} = \mathrm{CA}(\tilde{\mathbf{F}}_i, \mathbf{X}_i), \tag{8}$$

and then fuse them by a 1×1 convolution:

$$\mathbf{X} = \mathrm{Conv}_{1\times 1}\big(\mathrm{Concat}(\mathbf{Z}_{d\leftarrow e}, \mathbf{Z}_{e\leftarrow d})\big). \tag{9}$$

Here, $\mathrm{CA}(\mathrm{feature}_1, \mathrm{feature}_2)$ denotes a cross-attention operator that treats 'feature_1' as the query and 'feature_2' as the source of keys and values. In a standard implementation,

$$\mathrm{CA}(\mathbf{A}, \mathbf{B}) = \mathrm{softmax}\left(\frac{(W_q\mathbf{A})(W_k\mathbf{B})^\top}{\sqrt{d}}\right) W_v\mathbf{B}, \tag{10}$$

where W_q, W_k, and W_v are learnable projections, and d is the feature dimension. Thus, $\mathrm{CA}(\mathbf{X}_i, \tilde{\mathbf{F}}_i)$ lets each location in the decoder feature selectively aggregate complementary information from the encoder feature, while $\mathrm{CA}(\tilde{\mathbf{F}}_i, \mathbf{X}_i)$ performs the reverse aggregation. This bidirectional design encourages the decoder to focus on encoder regions that are most informative for lesion refinement and to suppress background responses persisting after encoding. The subsequent 1×1 convolution aligns the channel dimension and mixes the information coming from both directions.

As illustrated in Figure 2, the fused feature $\mathbf{X}$ is then refined by a four-branch multi-scale deformable unit. First, $\mathbf{X}$ is evenly split along the channel dimension:

$$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4 = \mathrm{Split}(\mathbf{X}), \tag{11}$$

where, if $\mathbf{X}$ has C channels, each sub-feature $\mathbf{x}_j$ contains $C/4$ channels. These four groups are processed in parallel by heterogeneous convolutions:

$$\mathbf{y}_1 = \mathrm{DWConv}_{1\times 1}(\mathbf{x}_1), \tag{12}$$

$$\mathbf{y}_2 = \mathrm{DeConv}^{(1)}_{3\times 3}(\mathbf{x}_2), \tag{13}$$

$$\mathbf{y}_3 = \mathrm{DeConv}^{(2)}_{3\times 3}(\mathbf{x}_3), \tag{14}$$

$$\mathbf{y}_4 = \mathrm{DeConv}^{(3)}_{3\times 3}(\mathbf{x}_4). \tag{15}$$

The first branch applies a depthwise 1×1 convolution to perform lightweight channel-wise filtering, while the remaining three branches adopt deformable convolutions with different dilation settings (or offset patterns). As a result, the three deformable branches possess different effective receptive fields and jointly form a multi-scale detail-enhancement module that is able to capture both

 doi:10.20944/preprints202603.0154.v1

thin boundaries and small lesion fragments. The outputs of all branches are concatenated and fused with the original decoder feature via a residual connection:

$$\mathbf{F}_i^{\text{mdcf}} = \mathbf{X}_i + \text{Concat}(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4). \tag{16}$$

In this formulation, $\mathbf{F}_i^{\text{mdcf}}$ denotes the output of the MDCF block; $\mathbf{X}_i$ is the input decoder feature; $\tilde{\mathbf{F}}_i$ is the SAFE-refined encoder feature; $\text{DWConv}_{1\times1}$ is a depthwise 1×1 convolution; $\text{DeConv}_{3\times3}^{(i)}$ is the i-th 3×3 deformable convolution branch (not a transposed convolution); and $\text{Split}(\cdot)$ and $\text{Concat}(\cdot)$ denote channel-wise splitting and concatenation, respectively.

The Upsampling block that follows each MDCF stage increases the spatial resolution of the decoder feature and prepares it for the next, finer decoding level. Concretely, the MDCF output $\mathbf{F}_i^{\text{mdcf}}$ is first passed through a 1×1 convolution to adjust the number of channels and reduce redundancy. The resulting feature map is then upsampled by a factor of 2, either using a learnable deconvolution or an interpolation-plus-convolution scheme, yielding the input feature for the subsequent decoding stage.

4. Results

4.1. Datasets

To comprehensively evaluate the performance of the proposed model, we conducted a series of experiments using two publicly accessible skin lesion segmentation datasets from ISIC (International Skin Imaging Collaboration), which originated from the "Skin Lesion Analysis Toward Melanoma Detection" challenge. These two datasets are ISIC-2016 [14] and ISIC-2018 [15], respectively. The images were meticulously annotated by professional experts with absolutely accurate labels. Specifically, 900 images were designated as training data, while 350 images were allocated for validation in the ISIC-2016 dataset [14]. The ISIC-2018 dataset [15] comprises 2,594 images, which were randomly divided into training, validation, and test sets at a ratio of 8:1:1.

4.2. Evaluation Metrics

To evaluate the effectiveness of our method in skin lesion segmentation, the primary evaluation metrics adopted in this paper include the Dice coefficient (also known as Dice Similarity Coefficient, DSC), Intersection over Union (IoU), Sensitivity, Specificity, and Accuracy. The Dice Coefficient is a statistical metric employed to measure the similarity between two sample sets, which has been extensively applied in medical image segmentation:

$$DC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{17}$$

The IoU metric quantifies the extent of overlap between the predicted segmentation region and the ground truth annotation:

$$IoU = \frac{TP}{TP + FP + FN} \tag{18}$$

The Accuracy (ACC) metric quantifies the proportion of correctly classified instances, encompassing both positive and negative predictions. The Sensitivity (SEN) assesses the proportion of actual positive cases that are correctly identified.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

$$SEN = \frac{TP}{TP + FN} \tag{20}$$

Where TP (True Positive) represents the number of correctly predicted positive cases, TN (True Negative) represents the number of correctly predicted negative cases, FP (False Positive) indicates the number of incorrectly predicted positive cases, and FN (False Negative) represents the number of

 doi:10.20944/preprints202603.0154.v1

incorrectly predicted negative cases. All metrics are dimensionless, providing a reliable evaluation of the model's segmentation performance.

4.3. Implementation Details

All experiments in this paper are conducted using the PyTorch 1.12.1 framework (with CUDA 11.3 support). The experiments are carried out on a server running Ubuntu 18.04.5 LTS, equipped with two Intel Xeon Gold 6230 CPUs , ten NVIDIA GeForce RTX 3090 GPUs, and a 2 TB solid-state drive.

For all experiments involving Our model, input images are resized to 352×352 pixels before being fed into the network. We employ the AdamW optimizer for model training, with the initial learning rate and weight decay both set to 1e-4. The network is trained for 200 epochs with a batch size of 8. A step-wise learning rate decay strategy is adopted, where the learning rate is multiplied by a factor of 0.1 every 200 epochs to ensure stable convergence.

To enhance the generalization capability of the model, data augmentation techniques such as random rotation and random flipping are utilized. The loss function is defined as a weighted combination of Binary Cross-Entropy (BCE) loss and Intersection over Union (IoU) loss (formulated here using Dice for structural supervision). The total objective function is defined as follows:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{CE} + \beta \cdot \mathcal{L}_{Dice} \tag{21}$$

where α and β represent the weighting coefficients. To determine the values of the weighting coefficients α and β, we performed a grid search by varying both parameters from 0 to 1 in increments of 0.1. The configuration yielding the best performance was α = 0.5 and β = 0.5, which was subsequently adopted for our experiments. To ensure a fair comparison with other baseline models, we followed their officially recommended training protocols and hyperparameters. The experiments primarily focus on the ISIC-2016 and ISIC-2018 datasets.

4.4. Ablation Study

To understand the importance of each component in our model, we tested different versions of our model on the ISIC 2016 dataset [14]. We measured performance using four metrics: Accuracy (ACC), Intersection over Union (IOU), Dice coefficient (DC), and Sensitivity (SEN). The results of the structural ablation experiments are presented in Table 1. The experimental results demonstrate that the Baseline model performs reasonably well in lesion detection, achieving an IOU of 86.9% and a DC of 82.6% on the ISIC 2016 dataset. However, its representation capability for lesion regions is insufficient. When we introduce PVT v2 as the backbone network, the model performance shows noticeable enhancement across all metrics. Specifically, the accuracy increases from 95.9% to 96.3%, the IOU improves from 86.9% to 87.0%, the Dice coefficient rises from 82.6% to 83.1%, and the sensitivity advances from 91.2% to 91.5%. These improvements indicate that the PVT v2 backbone network is more capable of extracting multi-scale features, thereby enhancing the model's representation ability for lesion regions.

Table 1. Result of an ablation investigation using the proposed method on the ISIC 2016 dataset.

Methods	ACC	IoU	DC	SEN
Baseline	0.959	0.869	0.926	0.912
PVT v2 backbone	0.963	0.870	0.931	0.915
PVT v2 backbone+SAFE	0.966	0.872	0.936	0.918
PVT v2 backbone+MDCF	0.968	0.869	0.938	0.920
PVT v2 backbone+SAFE+MDCF	0.970	0.873	0.941	0.921

Further incorporating the SAFE module on top of the PVT v2 backbone yields additional performance gains. The model achieves an accuracy of 96.6%, an IOU of 87.2%, a Dice coefficient of 93.6%, and a sensitivity of 91.8%. The SAFE module is specifically designed to improve encoder features at each resolution level by jointly leveraging both shallow and deep representations. By explicitly

 doi:10.20944/preprints202603.0154.v1

suppressing background responses while enhancing foreground lesion features, the SAFE module addresses one of the key challenges in skin lesion segmentation where background regions often exhibit similar visual characteristics to lesion areas. The integration of multi-level features enables the model to capture both fine-grained details from shallow layers and high-level semantic information from deep layers, resulting in more discriminative feature representations. This synergistic combination of shallow and deep features, coupled with explicit background suppression, leads to more precise localization of lesion boundaries and improved differentiation between lesion and non-lesion regions. As illustrated in Figure 3, the model equipped with SAFE demonstrates superior capability in capturing lesion regions within the images.

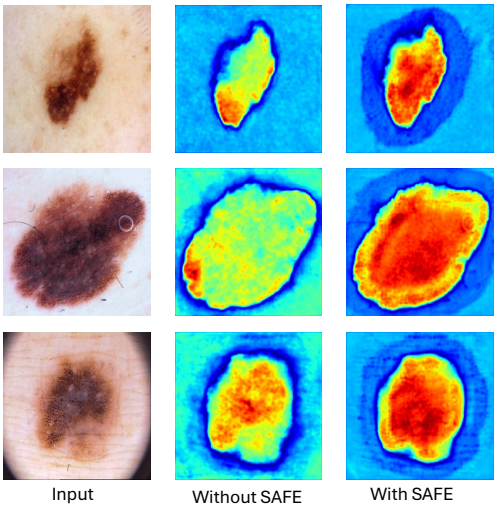

Figure 3. Visualization of decoder features for SAFE model. The deeper the red color, the higher the confidence that the region is a lesion area.

The complete model, which combines the PVT v2 backbone, SAFE module, and MDCF module, achieves the best performance across all evaluation metrics. It attains an accuracy of 97.0%, an IOU of 97.3%, a Dice coefficient of 94.1%, and a sensitivity of 92.1%. The MDCF module, through its multi-scale dense connection fusion mechanism, effectively integrates feature information from different hierarchical levels, further enhancing the model's capability to segment lesions with complex morphologies.

The ablation study conclusively demonstrates that each proposed component contributes meaningfully to the overall performance improvement. The progressive introduction of the PVT v2 backbone, SAFE module, and MDCF module results in incremental performance gains, validating the rationality of our design strategy. Moreover, the synergistic effects among these modules collectively drive the advancement of skin lesion segmentation performance, highlighting the effectiveness of our comprehensive architectural design.

4.5. Comparison with the State-of-the-Art Methods

To demonstrate the effectiveness of our proposed method, we conducted comprehensive comparative experiments against several state-of-the-art segmentation approaches on the ISIC-2016 [14] and ISIC-2018 [15] benchmark datasets. The compared methods include U-Net [10], Swin-Unet [22], TransUNet [12], H2Former [26], SUnet [27], and MSDUnet [28], which represent both classical CNN-based architectures and recent transformer-based models. All models were trained under identical experimental settings to ensure fair comparison, including the same data augmentation strategies, optimization parameters, and evaluation protocols.

 doi:10.20944/preprints202603.0154.v1

4.5.1. Quantitative Evaluation

Table 2 presents the quantitative comparison results on the ISIC-2016 dataset across four widely-used evaluation metrics: Accuracy (ACC), Intersection over Union (IOU), Dice Coefficient (DC), and Sensitivity (SEN). Our proposed method achieves the best performance across all metrics, with ACC of 0.970, IOU of 0.873, DC of 0.941, and SEN of 0.921. Compared to the second-best performing method MSDUnet, our approach demonstrates consistent improvements of 1.1% in ACC, 0.4% in IOU, 1.5% in DC, and 0.9% in SEN. When compared to the classical U-Net baseline, our method achieves substantial improvements of 6.8% in ACC, 7.0% in IOU, 5.9% in DC, and 3.5% in SEN, which clearly validates the effectiveness of our architectural innovations.

Table 2. A comparative study of different advanced segmentation techniques on the ISIC2016 dataset.

Methods	ACC	IoU	DC	SEN
U-Net [10]	0.902	0.803	0.882	0.886
Swin-Unet [22]	0.923	0.821	0.909	0.899
TransUNet [12]	0.932	0.833	0.913	0.901
H2Former [26]	0.956	0.852	0.918	0.906
SUnet [27]	0.954	0.866	0.922	0.909
MSDUnet [28]	0.959	0.869	0.926	0.912
Ours	**0.970**	**0.873**	**0.941**	**0.921**

The results on the ISIC-2018 dataset, summarized in Table 3, further confirm the superiority of our proposed method. Our approach consistently achieves the highest performance with ACC of 0.970, IOU of 0.836, DC of 0.902, and SEN of 0.928. The improvements over MSDUnet are 0.8% in ACC, 0.3% in IOU, 0.5% in DC, and 1.2% in SEN, while the gains over U-Net reach 4.9% in ACC, 3.0% in IOU, 4.4% in DC, and 2.9% in SEN. These consistent improvements across two different datasets demonstrate the strong generalization capability and robustness of our method.

Table 3. A comparative study of different advanced segmentation techniques on the ISIC2018 dataset.

Methods	ACC	IOU	DC	SEN
U-Net [10]	0.921	0.806	0.858	0.899
Swin-Unet [22]	0.933	0.815	0.862	0.906
TransUNet [12]	0.946	0.821	0.880	0.912
H2Former [26]	0.959	0.825	0.877	0.910
SUnet [27]	0.958	0.829	0.887	0.916
MSDUnet [28]	0.962	0.833	0.897	0.921
Ours	**0.970**	**0.836**	**0.902**	**0.928**

It is worth noting that transformer-based methods such as Swin-Unet, TransUNet, and H2Former generally outperform the classical U-Net, indicating the advantages of self-attention mechanisms in capturing long-range dependencies for medical image segmentation. However, our method surpasses these transformer-based approaches by effectively integrating multi-scale feature extraction with enhanced attention mechanisms, enabling more comprehensive representation learning and precise boundary delineation.

4.5.2. Qualitative Analysis

Beyond quantitative metrics, we provide qualitative visualization results to offer deeper insights into the segmentation performance of different methods. Figure 4 illustrates representative segmentation results on the ISIC-2016 dataset, comparing our method with TransUNet, SUnet, and MSDUnet alongside the ground truth annotations. The visual comparison reveals that our method produces segmentation masks with significantly clearer and more accurate boundaries, particularly for lesions with irregular shapes as shown in the second and third rows. In challenging cases where artifacts or noise are present, such as the example in the third row, our approach demonstrates superior robustness

 doi:10.20944/preprints202603.0154.v1

by generating clean segmentation masks that closely match the ground truth, while other methods tend to produce fragmented or incomplete predictions. Furthermore, for lesions with complex morphologies illustrated in the fourth row, our method better preserves the overall shape and structural integrity of the lesion regions, avoiding the over-segmentation or under-segmentation issues observed in competing approaches.

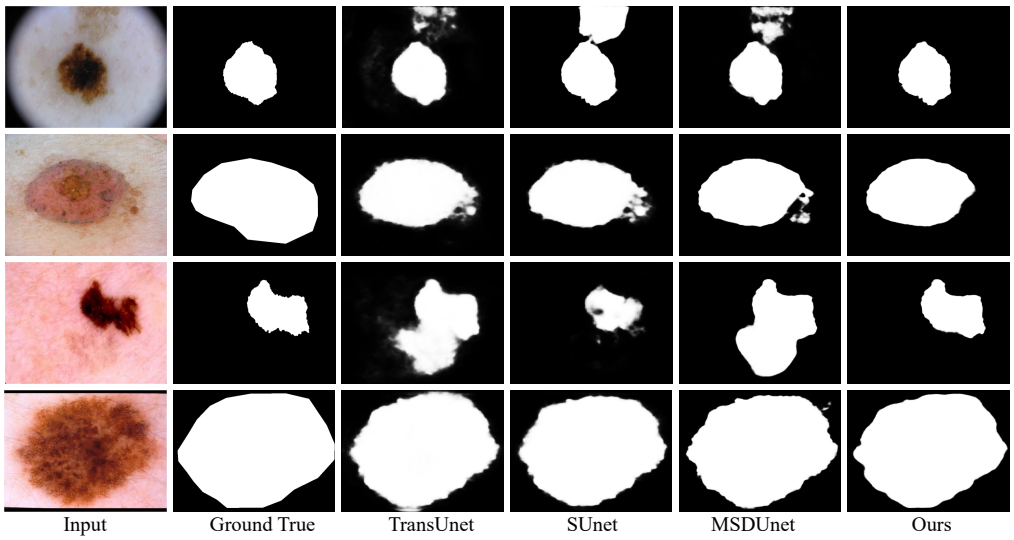

Figure 4. The qualitative analysis on the ISIC-2016 dataset.

Figure 5 presents the qualitative comparison on the ISIC-2018 dataset, which further validates the effectiveness of our proposed method across diverse lesion characteristics and imaging conditions. The second row demonstrates a particularly challenging case involving a small lesion, where our method successfully achieves precise segmentation while TransUNet exhibits significant under-segmentation, failing to capture the complete lesion extent. For lesions with highly irregular and fuzzy boundaries shown in the first and second rows, our approach produces more coherent and accurate segmentation masks that better align with the ground truth compared to competing methods, which often struggle with boundary ambiguity. Across all four representative cases spanning various lesion types, sizes, and imaging conditions, our method maintains consistent high-quality performance, demonstrating strong generalization capability and robustness to the inherent variability in dermoscopic images.

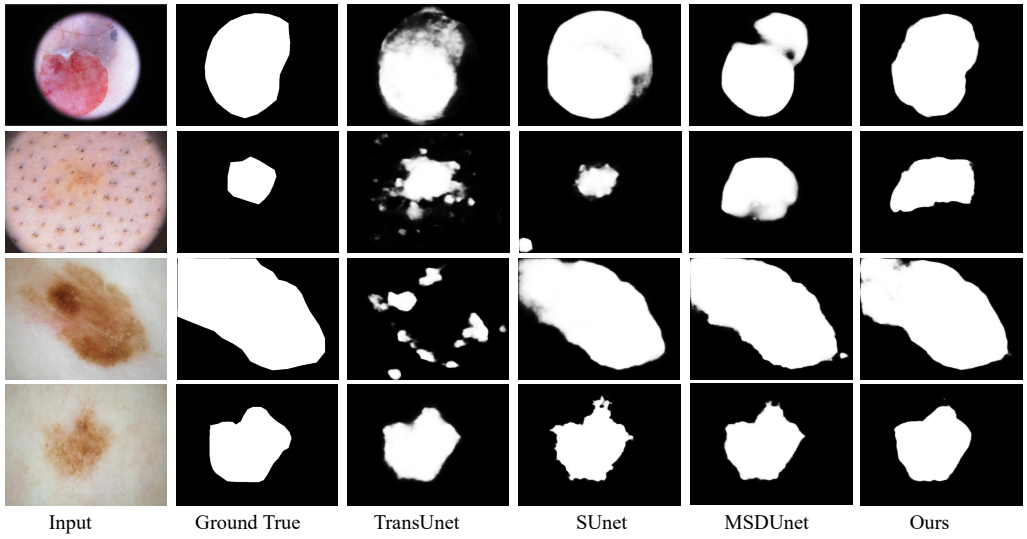

Figure 5. The qualitative analysis on the ISIC-2018 dataset.

 doi:10.20944/preprints202603.0154.v1

The superior performance of our proposed method can be attributed to the synergistic integration of the PVT v2 backbone with the SAFE and MDCF modules within our architecture. The pyramid structure of PVT v2 enables effective multi-scale feature extraction, capturing both fine-grained textural details and global contextual information, which is crucial for handling lesions of varying sizes and appearances. The SAFE module further refines these multi-scale features by bridging the feature discrepancy across network depths through its dual-branch architecture, where the base-feature branch preserves shallow-layer details while the offset-compensation branch adaptively aligns them with deep semantic information using dilated convolutions and global statistics. The MDCF module then establishes bidirectional cross-attention between decoder and SAFE-enhanced encoder features, facilitating precise localization of lesion boundaries by adaptively emphasizing discriminative features while suppressing irrelevant background information. Its four-branch multi-scale deformable convolution unit captures lesion boundaries and small fragments at heterogeneous receptive fields, which is particularly effective in challenging scenarios with low contrast, ambiguous edges, or irregular morphologies. This comprehensive architectural design enhances the model's ability to handle diverse lesion characteristics, including variations in size, shape, color, and texture patterns commonly observed in clinical dermoscopic images, ultimately achieving a superior balance between segmentation accuracy and computational efficiency.

In summary, both quantitative metrics and qualitative visualizations conclusively demonstrate that our proposed method establishes a new state-of-the-art for skin lesion segmentation on the ISIC-2016 and ISIC-2018 benchmark datasets, outperforming existing advanced approaches in terms of segmentation accuracy, boundary precision, and robustness to challenging imaging conditions.

5. Discussion

The experimental results presented in this study demonstrate that our proposed framework achieves state-of-the-art performance on both ISIC 2016 and ISIC 2018 datasets, validating the effectiveness of integrating PVT v2 backbone with SAFE and MDCF modules for skin lesion segmentation. The hierarchical pyramid structure of PVT v2 proves particularly effective for capturing lesion features at different scales, maintaining global contextual awareness while managing computational complexity through its spatial-reduction attention mechanism. This capability is especially valuable for dermoscopic images where lesions exhibit significant size variation, ranging from small melanomas to large nevi.

The consistent superior performance on both ISIC 2016 and ISIC 2018 datasets indicates strong generalization capability. Despite differences in image acquisition protocols, lesion types, and annotation quality between these datasets, our method maintains robust performance, suggesting that the learned representations capture fundamental characteristics of skin lesions rather than dataset-specific artifacts. While recent transformer-based approaches demonstrate advantages over classical CNN architectures, our method surpasses them by effectively combining the global modeling capability of transformers with targeted feature enhancement and fusion strategies. The performance gap suggests that architectural innovations beyond simply adopting self-attention mechanisms are necessary for optimal medical image segmentation. The qualitative results reveal that our method excels in challenging scenarios commonly encountered in clinical practice, including lesions with fuzzy boundaries, small lesions that are easily overlooked, and cases with imaging artifacts such as hair or air bubbles. The ability to produce clean, accurate segmentation masks in these difficult cases has direct implications for computer-aided diagnosis systems, potentially reducing false negatives in early melanoma detection and improving diagnostic confidence for dermatologists.

However, several limitations warrant discussion. While our method achieves superior segmentation accuracy, the integration of PVT v2 backbone, SAFE modules at multiple scales, and MDCF blocks with cross-attention mechanisms inevitably increases computational cost compared to lightweight models. Although we did not encounter memory constraints during training, deployment on resource-limited devices may require model compression techniques such as knowledge distillation, pruning, or quantization. Our evaluation is conducted on ISIC challenge datasets which, while widely used as

 doi:10.20944/preprints202603.0154.v1

benchmarks, primarily contain images acquired under controlled dermoscopic conditions that may not fully represent the variability encountered in real-world clinical practice. The datasets also exhibit class imbalance, with melanoma cases being relatively rare compared to benign lesions, and inter-annotator variability in ground truth labels may introduce noise affecting both training and evaluation.

The current framework provides deterministic predictions without quantifying uncertainty, which is crucial for clinical decision support. Knowing when the model is uncertain about its predictions can help clinicians identify cases requiring additional scrutiny or expert consultation. Additionally, our model relies solely on image data and does not leverage complementary information such as patient metadata or clinical history that dermatologists consider during diagnosis. While we provide feature visualization to demonstrate that SAFE enhances focus on lesion regions, the overall decision-making process remains largely opaque, and incorporating explainable AI techniques would increase transparency and build trust among medical practitioners. Furthermore, our method is specifically designed for melanoma and common skin lesion segmentation, and its performance on other dermatological conditions has not been assessed. The model also performs single-image segmentation without considering temporal information from follow-up images, which is important for monitoring lesion changes over time.

6. Conclusions

In this paper, we have presented a novel deep learning framework for automatic skin lesion segmentation that addresses key challenges in balancing segmentation accuracy and computational efficiency. By integrating the Pyramid Vision Transformer v2 backbone with Spatial-Aware Feature Enhancement and Multi-scale Dual Cross-attention Fusion modules, our method achieves state-of-the-art performance on the ISIC 2016 and ISIC 2018 benchmark datasets, with IOU scores of 87.33% and 83.67%, respectively.

The PVT v2 encoder leverages its hierarchical pyramid structure and spatial-reduction attention mechanism to extract multi-scale features efficiently, enabling the model to capture both fine-grained textural details and global contextual information essential for handling diverse morphological characteristics of skin lesions. The SAFE module refines encoder features at each resolution level through a dual-branch architecture that preserves shallow-layer spatial details while employing dilated convolutions to predict adaptive offsets that align shallow features with deep semantic representations, effectively bridging the feature discrepancy across network depths. The MDCF module enhances the decoder pathway through bidirectional cross-attention and multi-scale deformable convolutions, establishing dynamic feature fusion that emphasizes relevant lesion characteristics while filtering out redundant background information and capturing lesion boundaries at heterogeneous receptive fields.

Extensive experiments demonstrate that our method achieves superior performance compared to existing state-of-the-art approaches, with both quantitative metrics and qualitative visualizations confirming that our framework excels in challenging scenarios including lesions with irregular shapes, fuzzy boundaries, small sizes, and imaging artifacts. The ablation study validates the incremental contribution of each proposed component, with the complete model achieving the best performance across all evaluation metrics. The clinical implications of our work are significant, as accurate and efficient skin lesion segmentation is a critical step in computer-aided diagnosis systems for melanoma, where early detection can dramatically improve patient survival rates.

Future research directions include developing lightweight variants through model compression techniques to enable deployment on resource-constrained devices, incorporating uncertainty quantification to provide confidence estimates alongside predictions, integrating multi-modal information to enhance diagnostic accuracy, extending the framework to handle temporal analysis for lesion monitoring, exploring weakly-supervised learning to reduce annotation requirements, and validating performance on a broader spectrum of dermatological conditions. Our proposed framework represents a significant advancement in automatic skin lesion segmentation, offering a promising tool for im-

 doi:10.20944/preprints202603.0154.v1

proving early detection of skin cancer and demonstrating the potential of deep learning in addressing critical challenges in medical image analysis.

Author Contributions: Conceptualization, S.W.; Methodology, S.W.; Formal analysis, S.W. and P.Z.; Writing—original draft, S.W.; Writing—review & editing, S.W.; Visualization, P.Z.; Project administration, H.X.; Supervision, Z.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in "Skin Lesion Analysis Toward Melanoma Detection" challenge dataset, at arXiv:1605.01397. 2016.

Acknowledgments: This research was supported by Guangxi Regional Innovation Capacity Improvement Program under Grant No.Guike XT2503960034

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. R. L. Siegel *et al.*, "Cancer statistics, 2022," *CA Cancer J. Clin.*, vol. 72, no. 1, pp. 7–33, 2022.
2. H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021.
3. J. E. Gershenwald *et al.*, "Melanoma staging: Evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual," *CA Cancer J. Clin.*, vol. 67, no. 6, pp. 472–492, 2017.
4. H. Kittler *et al.*, "Diagnostic accuracy of dermoscopy," *Lancet Oncol.*, vol. 3, no. 3, pp. 159–165, 2002.
5. G. Argenziano *et al.*, "Dermoscopy of pigmented skin lesions: Results of a consensus net meeting conducted via the Internet," *J. Amer. Acad. Dermatol.*, vol. 48, no. 5, pp. 679–693, 2003.
6. C. Barata, M. E. Celebi, and J. S. Marques, "A survey of feature extraction in dermoscopy image analysis of skin cancer," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1096–1109, May 2019.
7. A. Lochan Sharma, K. Sharma, and P. Ghosal, "Skin lesion segmentation: A systematic review of computational techniques, tools, and future directions," *Comput. Biol. Med.*, vol. 196, 110842, 2025.
8. M. A. M. Almeida and I. A. X. Santos, "Classification models for skin tumor detection using texture analysis in medical images," *J. Imaging*, vol. 6, no. 6, p. 51, 2020.
9. A. Karimi, K. Faez, and S. Nazari, "DEU-Net: Dual-encoder U-Net for automated skin lesion segmentation," *IEEE Access*, vol. 11, pp. 134804–134821, 2023.
10. 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*, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds. Cham: Springer, 2015, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.
11. M. H. Jafari, N. Karimi, E. Nasr-Esfahani, S. M. R. Soroushmehr, S. Samavi, K. Ward, and K. Najarian, "Skin lesion segmentation in clinical images using deep learning," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, 2016, pp. 337–342, doi: 10.1109/ICPR.2016.7899656.
12. J. Chen, Y. Lu, Q. Yu, *et al.*, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
13. W. Wang, E. Xie, X. Li, *et al.*, "PVTv2: Improved Pyramid Vision Transformer for Dense Prediction," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 548–558. doi: 10.1109/CVPR52688.2022.00064.
14. D. Gutman *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2016 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)" arXiv:1605.01397, 2016.
15. N. C. Codella *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," arXiv:1902.03368, 2019.
16. V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.
17. A. Azhari, N. Yudistira, A. Widodo, et al., "Two-stage CNN with weakly supervised segmentation for skin lesion classification," *Multimedia Tools and Applications*, published online, 2025. doi: 10.1007/s11042-025-21091-8.

 doi:10.20944/preprints202603.0154.v1

18. F. Milletarì, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *Proc. Fourth International Conference on 3D Vision (3DV)*, Stanford, CA, USA, 2016, pp. 565-571. doi: 10.1109/3DV.2016.79.
19. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2999–3007. doi: 10.1109/ICCV.2017.324.
20. W. Wang, Y. Luo, and X. Wang, "BEFNet: A Hybrid CNN-Mamba Architecture for Accurate Skin Lesion Image Segmentation," in *Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Lisbon, Portugal, 2024, pp. 3795–3798. doi: 10.1109/BIBM62325.2024.10822106.
21. C. Yuan, D. Zhao, and S. S. Agaian, "UCM-Net: A lightweight and efficient solution for skin lesion segmentation using MLP and CNN," *Biomedical Signal Processing and Control*, vol. 96, p. 106573, 2024, doi: 10.1016/j.bspc.2024.106573.
22. H. Cao, Y. Wang, J. Chen, *et al.*, "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," *arXiv preprint arXiv:2105.05537*, 2021.
23. E. Xie, W. Wang, Z. Yu, *et al.*, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
24. T.-Y. Lin, P. Dollár, R. Girshick, K. He, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2117–2125. doi: 10.1109/CVPR.2017.106.
25. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Proc. European Conf. on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 801–818. doi: 10.1007/978-3-030-01234-2_49.
26. A. He, K. Wang, T. Li, C. Du, S. Xia, and H. Fu, "H2Former: An efficient hierarchical hybrid Transformer for medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 42, no. 9, pp. 2763–2775, 2023, doi: 10.1109/TMI.2023.3264513. :contentReference[oaicite:0]index=0
27. C.-M. Fan, T.-J. Liu, and K.-H. Liu, "SUNet: Swin Transformer with UNet for image denoising," in *Proc. 2022 IEEE Int. Symp. Circuits Syst. (ISCAS)*, pp. 2333–2337, 2022, doi: 10.1109/ISCAS48785.2022.9937486. :contentReference[oaicite:0]index=0
28. X. Li, L. Linli, X. Xing, H. Liao, W. Wang, Q. Dong, and C. Yuan, "MSDUNet: A model based on feature multi-scale and dual-input dynamic enhancement for skin lesion segmentation," *IEEE Trans. Med. Imaging*, vol. 44, no. 7, pp. 2819–2830, Jul. 2025, doi: 10.1109/TMI.2025.3549011. :contentReference[oaicite:0]index=0
29. A. Bilal, A. H. Khan, K. Almohammadi, S. A. A. Ghamdi, H. Long, and H. Malik, "PDCNET: Deep convolutional neural network for classification of periodontal disease using dental radiographs," *IEEE Access*, vol. 12, pp. 150147–150168, 2024.
30. L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2065–2074, Sep. 2017.
31. M. A. Khan, I. Haider, M. Nazir, A. Nabeeh, and E. S. M. El-Kenawy, "A multi-stage melanoma recognition framework with deep residual neural network and hyperparameter optimization-based decision support in dermoscopy images," *Expert Syst. Appl.*, vol. 215, Apr. 2023, Art. no. 119251.
32. G. Zhang, X. Shen, S. Chen, L. Liang, Y. Luo, J. Yu, and J. Lu, "DSM: A deep supervised multi-scale network learning for skin cancer segmentation," *IEEE Access*, vol. 7, pp. 140936–140945, 2019.
33. A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021. arXiv:2010.11929.
34. J. Valanarasu, R. Sindagi, V. Hacihaliloglu, and V. Patel, "Medical Transformer: Gated Axial-Attention for Medical Image Segmentation," in *Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021, pp. 36–46.
35. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2021, pp. 10012–10022.
36. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
37. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv preprint arXiv:1804.03999*. [Online]. Available: https://arxiv.org/abs/1804.03999

 doi:10.20944/preprints202603.0154.v1

38. S. Qamar, S. F. Qadri, R. Alroobaea, G. M. Alshmrani, and R. Jiang, "ScaleFusionNet: Transformer-guided multi-scale feature fusion for skin lesion segmentation," *Sci. Rep.*, vol. 15, Art. 34393, 2025, doi: 10.1038/s41598-025-17300-x. :contentReference[oaicite:3]index=3
39. W. Huang, X. Cai, Y. Yan, and Y. Kang, "MA-DenseUNet: A skin lesion segmentation method based on multi-scale attention and bidirectional LSTM," *Appl. Sci.*, vol. 15, no. 12, Art. 6538, Jun. 2025, doi: 10.3390/app15126538. :contentReference[oaicite:4]index=4
40. Y. Yin, J. Li, F. Zhang, et al., "Skin lesion segmentation with a multiscale input fusion U-Net incorporating Res2-SE and pyramid dilated convolution," *Sci. Rep.*, vol. 15, Art. 92447, 2025, doi: 10.1038/s41598-025-92447-1. :contentReference[oaicite:5]index=5
41. D. Li, X. Wu, and Q. Wei, "MSDTCN-Net: A multi-scale dual-encoder network for skin lesion segmentation," *Diagnostics*, vol. 15, no. 22, Art. 2924, Nov. 2025, doi: 10.3390/diagnostics15222924. :contentReference[oaicite:6]index=6
42. S. Shahin, X. Zhao, Y. … et al., "CTH-Net: A CNN and Transformer hybrid network for skin lesion segmentation," *iScience*, vol. 27, Art. 11042008, 2024, doi:10.1016/j.isci.2024.11042008. :contentReference[oaicite:8]index=8
43. Y. Li, T. Tian, J. Hu, and C. Yuan, "SUTrans-NET: A hybrid transformer approach to skin lesion segmentation," *PeerJ Comput. Sci.*, vol. 10, p. e1935, 2024, doi:10.7717/peerj-cs.1935. :contentReference[oaicite:9]index=9
44. M. Li, Y. Jiang, G. Cao, T. Xu, and R. Guo, "HyperFusionNet combines vision transformer for early melanoma detection and precise lesion segmentation," *Sci. Rep.*, vol. 15, Art. 30184, Nov. 2025, doi:10.1038/s41598-025-30184-1. :contentReference[oaicite:3]index=3
45. S. Perera, Y. Erzurumlu, D. Gulati, and A. Yilmaz, "MobileUNETR: A lightweight end-to-end hybrid vision transformer for efficient medical image segmentation," in *Proc. ECCV Workshops*, 2024. :contentReference[oaicite:11]index=11