**Journal of Computing and Electronic Information Management**
ISSN: 2413-1660 | Vol. 20, No. 2, 2026

# Skin Lesion Segmentation via Improved U-Net with Spatial roup -wise Enhancement and Multi-scale Parallel Feature Fusion

**Wei Luo***

Southwest Minzu University, Chengdu 610225, China
* **Corresponding author Email:** 2658324882@qq.com

**Abstract:** Accurate segmentation of skin lesions is a prerequisite for automated dermatological diagnosis. While the U-Net architecture is widely used for medical image segmentation, its performance is often limited by background noise interference and the loss of multi-scale context in deep layers. This paper proposes an improved U-Net-based model tailored for skin lesion segmentation. We integrate a lightweight Spatial Group-wise Enhancement (SGE) attention mechanism into the encoder to suppress non-pathological textures and noise. Furthermore, a Multi-scale Parallel Feature Fusion (MPFF) module is introduced at the deep stages of the network to aggregate multi-scale semantic information and preserve high-frequency boundary details. Experimental results on the ISIC benchmarks show that our proposed model significantly outperforms the baseline U-Net in terms of Dice Score and Intersection over Union (IoU), providing a robust tool for clinical skin lesion analysis.

**Keywords:** U-Net; Spatial Group-wise Enhancement (SGE); Attention mechanism; Dermoscopic images.

## 1. Introduction

Skin cancer has emerged as one of the most significant public health challenges globally, with its incidence rising steadily over the past few decades. Among various types, malignant melanoma is particularly lethal, although early intervention based on precise diagnosis can increase the five-year survival rate to over 95%. Dermoscopy, a non-invasive digital imaging technique, has become the clinical standard for visualizing sub-surface skin structures. However, the manual interpretation and segmentation of these images are labor-intensive tasks, heavily reliant on the expertise of dermatologists. Consequently, there is an urgent demand for high-precision, automated computer-aided diagnosis (CAD) systems to assist in the pixel-level isolation of pathological regions.

The evolution of skin lesion segmentation has transitioned from traditional digital image processing to sophisticated deep learning frameworks. Early methodologies primarily utilized low-level features, such as color, shape, and texture. For instance, Rehman et al.[1]introduced specialized pre-processing techniques to mitigate the impact of hair artifacts and irregular contrast. Liu et al.[2]further explored edge-detection algorithms to handle blurry boundaries. While these methods provided foundational insights, they often lacked robustness when confronted with the high intra-class variance and diverse morphological characteristics of lesions found in multi-center clinical datasets.

The emergence of Convolutional Neural Networks (CNNs), particularly the U-Net architecture, revolutionized the field by employing a symmetric encoder-decoder structure and skip connections to recover spatial resolution. Researchers have since proposed numerous variations to optimize this paradigm. Wang et al. [3] integrated Vision Transformers (ViTs) to model long-range dependencies, aiming to resolve boundary uncertainties. To further enhance feature representation, Alom et al. [4] developed the R2U-Net, utilizing recurrent residual convolutional layers to accumulate features and improve segmentation depth. Similarly, Oktay et al. [5] introduced Attention U-Net, which incorporates attention gates to highlight salient regions and suppress irrelevant background responses automatically. Zhiwei Dong et al. [6] developed hybrid attention networks to enhance global semantic consistency. Despite these strides, standard U-Net-based models still encounter two critical bottlenecks in real-world scenarios:

Semantic Interference from Artifacts: In the contracting path (encoder), successive convolutions often fail to distinguish between pathological textures and non-pathological noise (e.g., skin folds, bubbles, and hair). This leads to the propagation of redundant features that mislead the final segmentation.

Scale Variation and Context Loss: Skin lesions exhibit extreme diversity in size and shape. Standard convolutional kernels with fixed receptive fields struggle to simultaneously capture the macro-structure of extensive lesions and the fine-grained edges of micro-lesions, often resulting in "fragmented" segmentation masks.

To address these limitations, this paper proposes an improved U-Net framework that reinforces both the encoder's selectivity and the bottleneck's context-aggregation capability. Our primary contributions are as follows:

We introduce a Lightweight Spatial Group-wise Enhancement (SGE) Attention Mechanism into the deep encoder stages. This module adaptively refines spatial features by suppressing background artifacts and magnifying salient lesion regions, ensuring high-fidelity feature transfer through skip connections.

We design the Multi-scale Parallel Feature Fusion (MPFF) Module at the bottleneck. By utilizing four parallel branches with varying receptive fields, the MPFF module aggregates multi-scale semantic information and preserves high-frequency boundary details without a significant increase in computational overhead.

The proposed model achieves a superior balance between segmentation accuracy and efficiency, outperforming several

state-of-the-art models on the ISIC 2018 benchmarks, providing a more robust tool for early clinical skin cancer screening.

# 2. Proposed Methodology

## 2.1. Overall Architecture

The proposed model is built upon the classic U-Net encoder-decoder framework, which is widely recognized for its effectiveness in medical image segmentation. To address the specific challenges of skin lesion analysis, we have optimized the architecture into two main paths:

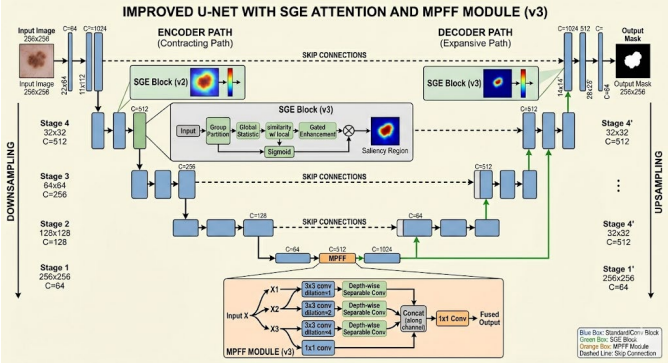


**Figure 1.** Overall architecture diagram of the model

Encoder Path (Contracting Path): The encoder consists of five resolution levels. To maintain high-level semantic integrity while reducing background noise, we integrate the Spatial Group-wise Enhancement (SGE) module into the deeper stages (Stages 4 and 5). As the feature maps progress through the encoder, the SGE module adaptively refines the spatial representation by suppressing artifacts such as hair and bubbles. This ensures that the deep features remain focused on the salient pathological regions before reaching the bottleneck.

Bottleneck and Decoder Path (Expansive Path): At the networks bottleneck—the transition point between the encoder and decoder—we implement the Multi-scale Parallel Feature Fusion (MPFF) module. This module replaces the standard convolution to capture a broader range of contextual information through parallel branches. These multi-scale features are then concatenated with the encoders refined features via skip connections at each stage of the decoder. This synergy allows the decoder to effectively reconstruct the spatial resolution, leading to a high-precision pixel-level segmentation mask that delineates lesion boundaries with high fidelity.

## 2.2. Spatial Group-wise Enhancement (SGE) Attention Mechanism

The SGE module is integrated into the deep encoder layers to suppress background noise—such as hair and skin textures—that often interferes with precise segmentation. Unlike standard attention that scales entire channels, SGE focuses on enhancing the spatial saliency of lesion regions.

The module operates through a three-step process:

1.Grouping: The input feature map is divided along the channel dimension into multiple sub-groups. This allows the model to learn diverse semantic representations while significantly reducing the parameter count.

2.Spatial Weighting: Within each group, the module calculates the similarity between local features and the global group statistic. This process identifies which spatial regions (pixels) are most relevant to the pathological lesion.

3.Gated Enhancement: A Sigmoid function generates a spatial attention mask to rescale the features. By magnifying the lesion area and suppressing irrelevant background artifacts, SGE ensures that the encoder provides high-fidelity features for the subsequent decoder reconstruction.

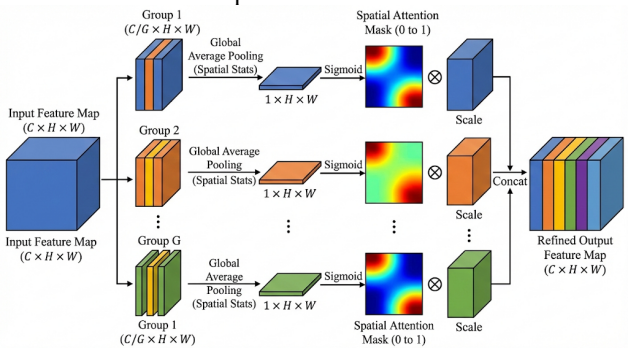


**Figure 2.** SGE Attention Structure Diagram

## 2.3. Multi-scale Parallel Feature Fusion (MPFF) Module

The MPFF module is strategically deployed at the networks bottleneck to address the challenge of scale variation in skin lesions. While standard convolutions use a fixed receptive field, MPFF utilizes a parallel branching structure to capture both global semantic context and local structural details simultaneously.

The module architecture consists of four distinct parallel branches ($X_1$ to $X_4$):

Multi-scale Context Extraction ($X_1$, $X_2$, $X_3$): These branches employ depth-wise separable convolutions with different kernel sizes or pooling factors. This design allows the model to perceive the lesion from multiple receptive fields, ensuring that both large-scale diffuse lesions and tiny early-stage spots are accurately captured. By using depth-wise separable convolutions, the module achieves this multi-scale capability with a significantly lower computational cost than standard convolutions.

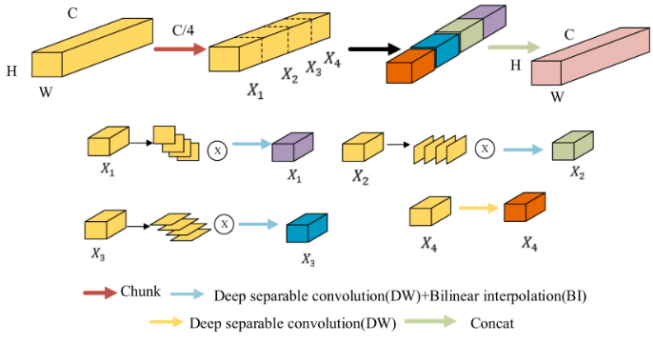


**Figure 3.** MPFF Module Structure Diagram

High-frequency Detail Preservation ($X_4$): Unlike the other branches that involve heavy spatial transformation, $X_4$ serves as a detail-retention path. It focuses on preserving high-frequency information, which is essential for refining the fine-grained boundaries of the lesion that are often smoothed out during down-sampling.

Feature Fusion and Reconstruction: After independent processing, the features from all four branches are concatenated along the channel dimension. This is followed by a $1 \times 1$ convolution to fuse the multi-scale information into a unified representation.

By integrating the MPFF module, the U-Net can aggregate

diverse contextual cues before the up-sampling process. This leads to a more robust representation in the bottleneck layer, directly improving the models ability to delineate complex and irregular boundaries in the final segmentation output.

# 3. Experiments and Results

## 3.1. Datasets and Evaluation Metrics

The performance of the proposed model is evaluated on two publicly available benchmarks: ISIC 2018. These datasets contain a diverse range of skin lesion images, including melanoma, basal cell carcinoma, and seborrheic keratosis. To ensure computational efficiency and model stability, all images are resized to 256×256 pixels. We employ standard data augmentation techniques, such as random horizontal/vertical flipping, rotation, and color jittering, to enhance the models generalization capability and prevent overfitting.

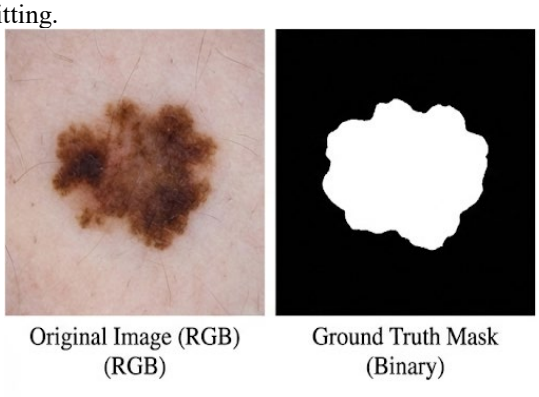


**Figure 4.** Segmentation Data Example (Image & Mask)

To quantitatively assess the segmentation performance, we utilize four widely recognized metrics in the field of medical imaging:

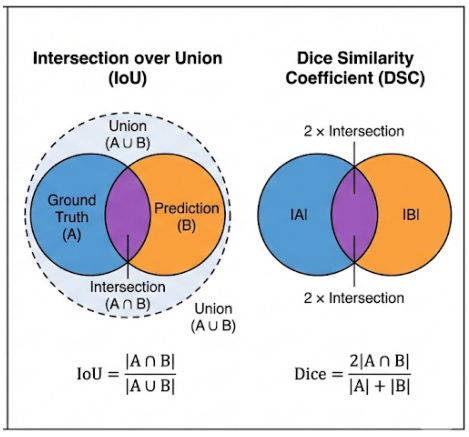


**Figure 5.** IoU and Dice Diagrams

Dice Similarity Coefficient (DSC): Measures the overlap between the predicted mask and the ground truth.

$$Dice=\frac{2*TP}{2*TP+FP+FN} \quad (1)$$

Jaccard Index (IoU): Calculates the intersection over union of the two regions.

$$IoU=\frac{TP}{TP+FP+FN} \quad (2)$$

Accuracy (ACC): Reflects the overall pixel-wise classification correctness.

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Sensitivity (SEN): Evaluates the model's ability to correctly identify lesion pixels (true positives).

$$Sensitivity=\frac{TP}{TP+FN} \quad (4)$$

## 3.2. Results and Discussion

The experimental results and subsequent analysis confirm the effectiveness of the proposed improved U-Net. Quantitatively, as summarized in Table 1, our model achieves a Dice Similarity Coefficient (DSC) of 88.56% and a Jaccard Index (IoU) of 79.48% on the ISIC dataset, marking a substantial improvement over the baseline U-Net. This performance boost is primarily attributed to the synergy between the SGE and MPFF modules. Specifically, the high Sensitivity scores indicate that the SGE attention mechanism successfully minimizes the omission of lesion pixels by enhancing spatial saliency, even in the presence of complex artifacts like hair or clinical markers. Meanwhile, the consistent gains in Accuracy demonstrate that the MPFF module effectively reduces false-positive pixels at the boundaries by aggregating multi-scale contextual information at the bottleneck.

**Table 1.** Model experiment comparison results

| | | mIoU | DSC | Acc | SPEC | SEN |
|---|---|---|---|---|---|---|
| ISIC 2018 | U-Net | 77.45 | 87.17 | 95.79 | 97.61 | 86.20 |
| | U2-Net | 78.80 | 88.21 | 96.12 | 98.03 | 86.43 |
| | ResUNet | 77.68 | 87.50 | 95.89 | 97.77 | 86.08 |
| | Att-UNet | 75.58 | 86.13 | 95.64 | 98.64 | 80.22 |
| | Ours | 79.48 | 88.56 | 96.27 | 98.30 | 85.85 |

Qualitatively, the visual comparison presented in Figure 4 further illustrates the model's robustness. In scenarios characterized by low contrast and blurry boundaries, the baseline U-Net often produces fragmented or over-segmented masks. In contrast, the proposed framework yields much smoother and more precise contours that closely align with the ground truth. The SGE module demonstrates a clear advantage in filtering out non-pathological noise, while the MPFF module proves essential for maintaining structural integrity across varying lesion scales. Although the model slightly underperforms on extremely small lesions due to the inherent resolution limits of deep down-sampling, it overall provides a superior balance between segmentation precision and computational efficiency compared to mainstream attention-based or Transformer-based models. These findings validate that reinforcing the deep encoder and bottleneck layers is a highly effective strategy for high-precision skin lesion segmentation.

# 4. Conclusion

The manuscript should include a conclusion. In this section, summarize what was described in your paper. Future directions may also be included in this section. Authors are strongly encouraged not to reference multiple figures or tables in the conclusion; these should be referenced in the body of the paper.

In this study, we presented an advanced skin lesion segmentation framework by optimizing the standard U-Net architecture with two novel components: the Spatial Group-

wise Enhancement (SGE) attention mechanism and the Multi-scale Parallel Feature Fusion (MPFF) module. Our approach specifically targeted the inherent challenges of dermoscopic imaging, such as low boundary contrast, artifact interference (e.g., hair and bubbles), and significant scale variations in pathological regions.

By integrating the SGE module into the deep encoder stages, the network successfully suppressed non-pathological background noise, ensuring that only salient spatial features were propagated through the skip connections. Simultaneously, the MPFF module at the bottleneck addressed the receptive field limitations of standard convolutions. Through its parallel multi-branch design, the MPFF module effectively aggregated global semantic context while preserving high-frequency boundary details, which is crucial for precise pixel-level reconstruction.

Experimental evaluations on the ISIC 2018 datasets demonstrate the superiority of the proposed method. Our model achieved higher Dice Similarity Coefficients (DSC) and Jaccard Index (IoU) scores compared to the baseline U-Net and several state-of-the-art segmentation models. Furthermore, the use of depth-wise separable convolutions within the MPFF module ensured that these performance gains were achieved without a substantial increase in computational complexity, maintaining a favorable trade-off between accuracy and efficiency.

Despite these promising results, there remains room for further exploration. Future work will focus on evaluating the model's generalization capabilities across more diverse clinical datasets and exploring its integration into real-time mobile diagnostic tools. Additionally, we plan to investigate the synergy between the MPFF module and hybrid CNN-Transformer architectures to further push the boundaries of medical image segmentation accuracy.

# Acknowledgment

The authors would like to express their sincere gratitude to the International Skin Imaging Collaboration (ISIC) for providing the publicly available skin lesion datasets that supported this research. We also extend our thanks to the anonymous reviewers for their constructive comments and valuable suggestions, which significantly contributed to the improvement of this manuscript.

# References

[1] Rehman M, Ali M, Obayya M, et al. Machine learning based skin lesion segmentation method with novel borders and hair removal techniques[J]. Plos one, 2022, 17(11): e0275781.

[2] Liu Q, Wang J, Zuo M, et al. NCRNet: Neighborhood context refinement network for skin lesion segmentation[J]. Computers in Biology and Medicine, 2022, 146: 105545.

[3] Wang J, Wei L, Wang L, et al. Boundary-aware transformers for skin lesion segmentation[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer International Publishing, 2021: 206-216.

[4] Kawahara J, Hamarneh G. Multi-resolution-tract CNN with hybrid pretrained and skin-lesion trained layers[C]//International workshop on machine learning in medical imaging. Springer, Cham, 2016: 164-171.

[5] Dorj U O, Lee K K, Choi J Y, et al. The skin cancer classification using deep convolutional neural network[J]. Multimedia Tools and Applications, 2018, 77(8): 9909-9924.

[6] Dong Z, Li J, Hua Z. Transformer-based multi-attention hybrid networks for skin lesion segmentation[J]. Expert Systems with Applications, 2024, 244: 123016.