AULUNet: An Adaptive Ultra-Lightweight U-Net Framework for Efficient Skin Lesion Segmentation in Resource-Constrained Environments

Md Maklachur Rahman[1]
maklachur@tamu.edu
Soon Ki Jung[2]
skjung@knu.ac.kr
Tracy Hammond[1]
hammond@tamu.edu

[1] Department of Computer Science & Engineering, Texas A&M University, College Station, USA

[2] School of Computer Science & Engineering, Kyungpook National University, Daegu, South Korea

Abstract

High accuracy in skin lesion segmentation is crucial for mobile and embedded medical applications, where most existing models are too computationally demanding. We introduce AULUNet, an ultra-lightweight U-Net framework that addresses this limitation through three key innovations: Adaptive Kernel Fusion (AKF) for balanced extraction of local and global features, Residual Global Fusion (RGF) for refining bottleneck representations with global context, and Lightweight Skip Gate (LSG) for selective skip connection enhancement. On ISIC2017, AULUNet achieves 85.32 % mIoU and 92.08 % DSC, surpassing standard U-Net by 5.77 % and 3.47 %, respectively. On ISIC2018, it achieves 82.44 % mIoU and 90.37 % DSC, delivering gains of 7.80 % and 4.89 % over the baseline. Notably, our model requires only 0.029 M parameters and 0.069 GFLOPs, reducing parameter count by 99.63 % (from 7.773 M to 0.029 M) and compute by 99.50 % (from 13.758 GFLOPs to 0.069 GFLOPs), while maintaining consistent performance across diverse lesion structures. We validate the robustness of our design through extensive experiments and ablation studies. Our code is publicly available at `https://github.com/maklachur/AULUNet`.

1 Introduction

Medical image segmentation in resource-constrained environments presents a significant challenge, as it requires balancing high segmentation accuracy with strict computational efficiency constraints [22, 23, 29, 32]. While deep learning has significantly advanced segmentation performance, many existing models suffer from high computational and memory requirements, making them unsuitable for mobile and low-power devices [4, 20, 33, 34].

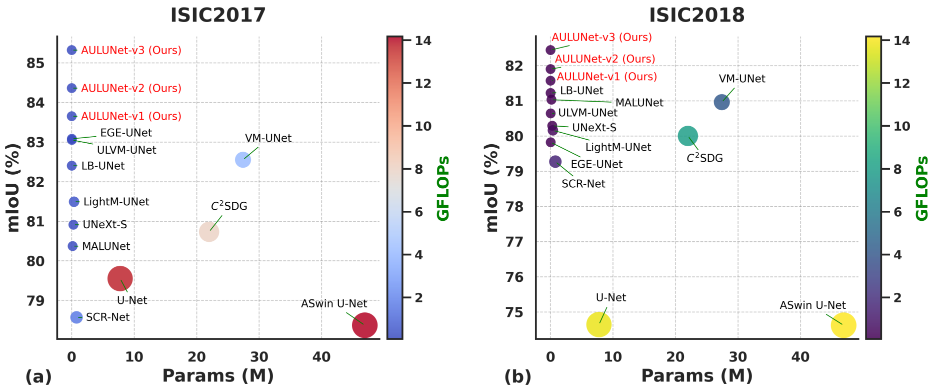

Figure 1: Comparative visualization of results on the ISIC2017 and ISIC2018 datasets, shown in (a) and (b), respectively. The X-axis indicates the number of parameters in millions (M), the Y-axis denotes mIoU (%), and the color gradient represents GFLOPs.

Skin lesion segmentation is critical for the early detection of malignant melanoma, which can greatly reduce mortality rates [9]. However, variations in lesion size, shape, texture, and artifacts (e.g., hair, low contrast, imaging inconsistencies) render this task complex [9]. The key challenge is to achieve high segmentation accuracy while minimizing computational costs for real-time mobile health applications [22, 23, 27, 29].

UNet [20] and its variants [3, 12, 18, 28, 34] are widely used in medical image segmentation due to their encoder-decoder architecture and multi-scale feature extraction via skip connections. For instance, UNet++ [34] improves feature fusion with redesigned skip pathways and Attention UNet [18] enhances focus on relevant regions using attention mechanisms. Additionally, ResUNet++ [14] leverages residual learning and atrous convolutions to strengthen feature representation and MultiResUNet [13] employs multi-resolution encoding for improved performance. More recently, transformer-based models such as TransUNet [5], Swin-UNet [4], and TransFuse [33] have achieved state-of-the-art performance by capturing long-range dependencies. Despite these advances, such models remain computationally expensive and are less practical in resource-constrained settings [22, 23].

To tackle these challenges, lightweight architectures such as UNeXt [24], MALUNet [22], EGE-UNet [23], LB-UNet [29], and UltraLight VM-UNet [27] have emerged. UNeXt [24] leverages tokenized MLPs to reduce model complexity but often struggles to capture fine-grained details. In contrast, MALUNet [22] and EGE-UNet [23] incorporate attention mechanisms to better balance parameter efficiency with segmentation accuracy. LB-UNet [29] refines boundaries for precision, and UltraLight VM-UNet [27] employs Mamba-based designs [19] to lower computational cost. Yet, these models often sacrifice accuracy for efficiency, reducing their practicality in real-world mobile health applications.

In this paper, we propose AULUNet, an ultra-lightweight UNet for skin lesion segmentation in resource-constrained environments. Our model achieves high accuracy with minimal computational overhead, making it ideal for real-time mobile applications. A visual comparison with other models is shown in Figure 1. AULUNet features three key modules. The Adaptive Kernel Fusion (AKF) module enhances multi-scale feature extraction by processing input features through two parallel depthwise convolution branches. A local path with a 3×3 kernel ($d = 1$) captures fine details, while a dilated path with a 3×3 kernel ($d = 2$) enlarges the receptive field. An adaptive gating mechanism, combining global average pooling, a 1×1 convolution, and sigmoid activation, dynamically fuses these features. The Residual Global Fusion (RGF) module refines bottleneck features by computing a global descriptor via adaptive average pooling, modulating it with a 1×1 convolution and tanh activation, and fusing it residually with the original feature map. The Lightweight Skip Gate (LSG) module improves encoder-decoder fusion by ensuring channel consistency with a 1×1 convolution and selectively propagating informative features through a gating mechanism based on 1×1 convolution, batch normalization, and sigmoid activation.

Our main contributions are summarized as follows:

- We propose AULUNet, a novel ultra-lightweight architecture for skin lesion segmentation that achieves high performance with only 0.029M parameters and 0.069 GFLOPs.
- We introduce three modules, including AKF for enhanced multi-scale feature extraction, RGF for refined global context integration, and LSG for improved skip connection fusion.
- Extensive experiments on the ISIC2017 and ISIC2018 datasets show that AULUNet consistently outperforms existing lightweight models in terms of Dice Similarity Coefficient (DSC), mean Intersection over Union (mIoU), and computational efficiency.

2 Methodology

Our AULUNet (Figure 2) uses a UNet-based encoder-decoder with six stages. In the encoder, an initial convolution at full resolution is immediately followed by the AKF module to learn multi-scale features. At each stage, average pooling reduces the spatial resolution while pointwise convolutions increase the channel dimensions. In the bottleneck, the RGF module refines the features by integrating global context via adaptive pooling and a 1×1 convolution with tanh activation, combined with residual fusion. The decoder gradually restores resolution using bilinear interpolation and pointwise convolutions. At every upsampling step, the LSG module fuses the encoder and decoder features through a gating mechanism, ensuring that only the salient information is propagated. Finally, a 1×1 convolution followed by sigmoid activation produces the segmentation output.

2.1 Adaptive Kernel Fusion (AKF)

Inspired by [15, 25], we designed the AKF mechanism based on the observation that skin lesions vary greatly in size. To efficiently capture multi-scale features, our method dynamically integrates fine-grained local features and global contextual understanding. Given an input feature map $\mathbf{x} \in \mathbb{R}^{B \times C \times H \times W}$, we process $\mathbf{x}$ through two parallel depthwise convolutional paths. The local path $\mathbf{F}_l$ applies a 3×3 depthwise convolution with dilation $d = 1$ to capture fine details, while the dilated path $\mathbf{F}_g$ employs a 3×3 depthwise convolution with dilation $d = 2$ to capture broader context. We empirically select $d = 1$ and $d = 2$ (see Table 4) as they provide a balanced trade-off: $d = 1$ preserves local details critical for small lesions and d=2 enlarges the receptive field sufficiently to detect larger, diffuse lesions without incurring excessive computational cost [6, 10, 30].

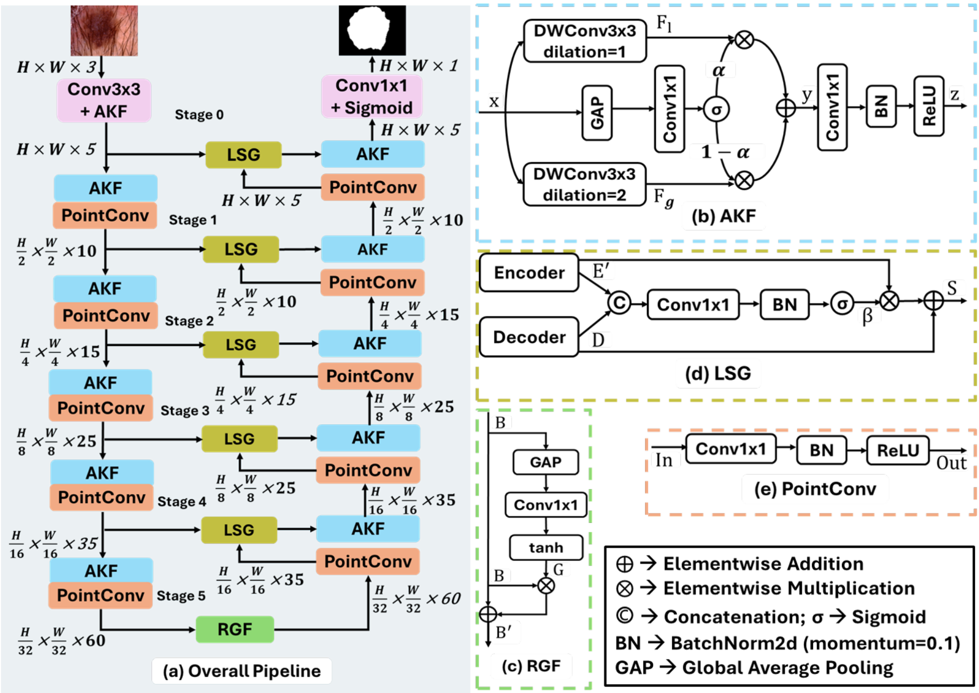

Figure 2: Overview of our proposed AULUNet. (a) The overall pipeline follows an encoder-decoder design with six stages, where each encoder–decoder stage employs an Adaptive Kernel Fusion (AKF) and pointwise convolutions, with Residual Global Fusion (RGF) at the bottleneck. Skip connections are integrated to refine the encoder-decoder feature via Lightweight Skip Gate (LSG) modules, (b) Detailed structure of AKF, (c) RGF module, (d) LSG module, and (e) PointConv block.

To adaptively fuse these features, we first derive a global descriptor via global average pooling (GAP), processed through a 1×1 convolution and sigmoid activation to yield dynamic gating weights $\alpha \in \mathbb{R}^{B \times C \times 1 \times 1}$, which control the balance between local and global contributions:

$$\alpha = \sigma\left(\mathrm{Conv}_{1\times 1}(\mathrm{GAP}(\mathbf{x}))\right). \tag{1}$$

We then computed the enhanced feature by weighting these paths with α, followed by a 1×1 convolution, batch normalization (momentum=0.1), and ReLU activation to produce the output $\mathbf{z}$:

$$\mathbf{z} = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times 1}(\alpha \odot \mathbf{F}_l \oplus (1-\alpha) \odot \mathbf{F}_g)\right)\right). \tag{2}$$

This output $\mathbf{z}$ is then passed to the next stage for further processing.

2.2 Residual Global Fusion (RGF)

We introduce an RGF module to integrate global context directly into the bottleneck features and enhance long-range dependency modeling. For a bottleneck feature map $\mathbf{B} \in \mathbb{R}^{B \times C \times H \times W}$, we first compute a global descriptor via GAP and then transform it using a 1×1

convolution followed by a tanh activation to obtain modulation coefficients $\mathbf{G} \in \mathbb{R}^{B\times C\times 1\times 1}$ with values in $[-1, 1]$:

$$\mathbf{G} = \tanh\Big(\mathrm{Conv}_{1\times 1}(\mathrm{GAP}(\mathbf{B}))\Big), \tag{3}$$

$$\mathbf{B}' = \mathbf{B} \oplus (\mathbf{B} \odot \mathbf{G}). \tag{4}$$

where $\oplus$ and $\odot$ denote element-wise addition and multiplication, respectively. This mechanism selectively enhances or suppresses channel responses based on the global context, refining the bottleneck features and capturing non-linear global interactions without introducing significant computational overhead.

2.3 Lightweight Skip Gate (LSG)

Traditional skip connections forward all encoder features to the decoder, which can introduce redundant or irrelevant information [9, 34]. Moreover, many existing methods rely on complex gating mechanisms that are computationally expensive [3, 9]. In contrast, our LSG module is designed to be lightweight while selectively transferring the salient and refined encoder features. We first ensure channel consistency via a 1×1 convolution when needed on a given decoder feature map $\mathbf{D} \in \mathbb{R}^{B\times C_D\times H\times W}$ and an encoder feature map $\mathbf{E} \in \mathbb{R}^{B\times C_E\times H\times W}$:

$$\mathbf{E}' = \begin{cases} \mathbf{E}, & \text{if } C_D = C_E, \\ \mathrm{Conv}_{1\times 1}(\mathbf{E}), & \text{otherwise.} \end{cases} \tag{5}$$

The transformed encoder feature $\mathbf{E}'$ is concatenated with $\mathbf{D}$ along the channel dimension. This concatenated feature is then processed through a 1×1 convolution, batch normalization, and sigmoid activation to generate the attention weights $\beta \in [0, 1]^{B\times C_D\times H\times W}$, which are subsequently used to fuse the encoder and decoder features via element-wise addition. We define these operations as:

$$\beta = \sigma\Big(\mathrm{BN}\Big(\mathrm{Conv}_{1\times 1}([\mathbf{D}, \mathbf{E}'])\Big)\Big) \tag{6}$$

$$\mathbf{S} = \mathbf{D} \oplus \beta \odot \mathbf{E}', \tag{7}$$

where $\odot$ and $\oplus$ denote element-wise multiplication and addition, respectively. This lightweight gating mechanism reduces redundancy by ensuring that only the most informative encoder features are forwarded to the decoder without integrating complex and heavy gating mechanisms.

3 Experiments and Results

3.1 Datasets, Implementation Details, and Evaluation

We evaluate our model on two widely used skin lesion segmentation datasets: ISIC2017 [8] and ISIC2018 [7], provided by the International Skin Imaging Collaboration (ISIC). Following previous works [21, 27, 31, 32], we preprocess the training, validation, and test sets by resizing all images to 256×256 pixels and applying normalization. We employ standard data augmentation techniques, including horizontal and vertical flipping and random rotations. Comparative evaluations are performed on both datasets, and ablation studies are conducted on ISIC2018 to analyze the contributions of each module.

All experiments are conducted on a single NVIDIA GeForce RTX 3090 Ti GPU with 24GB VRAM. We train our model using a combination of binary cross-entropy and Dice loss [2] to balance per-pixel accuracy and overlap-based segmentation quality:

$$\begin{aligned}\mathcal{L}_{\text{BCE}} &= -\frac{1}{M}\sum_{j=1}^{M}\left[y_j \log(\hat{y}_j) + (1-y_j)\log(1-\hat{y}_j)\right],\\ \mathcal{L}_{\text{Dice}} &= 1 - \frac{2|A\cap B|}{|A|+|B|}, \quad \mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}},\end{aligned} \tag{8}$$

where M is the number of pixels in each image, y_j and $\hat{y}_j$ denote the ground truth label and predicted probability for pixel j, and $|A|$, $|B|$ represent the sizes of the ground truth and predicted masks, respectively.

We train for 400 epochs using the AdamW optimizer[17] with a batch size of 8, and employ a cosine-annealing learning-rate scheduler [16] that starts at 0.001 and decays to a minimum of 0.00001.

For evaluation, we report the mean Intersection over Union (mIoU) and Dice Similarity Coefficient (DSC) as our primary metrics. Additionally, we quantify model complexity using the number of parameters (in millions) denoted as Params (M) and the number of GFLOPs computed for an input size of 256×256.

3.2 Model Variants and Trade-offs

We evaluated three AULUNet configurations to balance segmentation accuracy with computational efficiency (see Figure 1). Our baseline, AULUNet-v1, employs a six-stage encoder-decoder with channel dimensions $[4,8,16,24,36,60]$. We then increased the capacity of the mid layers in AULUNet-v2 by adjusting the channels to $[4,8,16,24,40,60]$, which provided insight into the role of mid-level features. Our final variant, AULUNet-v3, uses a revised configuration $[5,10,15,25,35,60]$ that achieves the best balance between model complexity and segmentation performance. These results demonstrate that even subtle channel adjustments can lead to different results and complexity, thereby guiding us to an optimal design for resource-constrained applications.

3.3 Comparison with SOTA Methods

We evaluate the performance of our proposed AULUNet variants against a wide range of SOTA segmentation models on the ISIC2017 [8] and ISIC2018 [7] datasets, as shown in Table 1. For a fair comparison, we reproduce all compared results on the same train, validation, and test splits using their official implementations. Our evaluation includes traditional models such as U-Net [20] and ATTENTION SWIN U-NET (ASwin U-Net) [1], along with recent lightweight architectures including UNeXt-S [24], MALUNet [22], EGE-UNet [23], LB-UNet [29], and UltraLight VM-UNet (ULVM-UNet) [27].

On ISIC2017, our AULUNet-v3 achieves 85.32% mIoU and 92.08% DSC, outperforming the strongest baseline, EGE-UNet, by 2.24% in mIoU and 1.32% in DSC. While EGE-UNet uses 0.053 million parameters and 0.072 GFLOPs, our model reduces these to 0.029 million and 0.069 GFLOPs, yielding 45.3% fewer parameters and 4.17% lower compute. Compared to LB-UNet, AULUNet-v3 offers 2.92% higher mIoU and 1.73% higher DSC

Dataset	Model	Params (M)↓	GFLOPs↓	mIoU (%)↑	DSC (%)↑	Year
ISIC2017	U-Net [20]	7.773	13.758	79.55	88.61	MICCAI15
	SCR-Net [26]	0.801	1.567	78.57	88.00	AAAI21
	ASwin U-Net [1]	46.910	14.181	78.37	87.87	ISBI23
	C^2SDG [11]	22.001	7.972	80.73	89.34	MICCAI23
	UNeXt-S [24]	0.302	0.103	80.91	89.45	MICCAI22
	MALUNet [22]	0.175	0.083	80.37	89.11	BIBM22
	EGE-UNet [23]	0.053	0.072	83.08	90.76	MICCAI23
	VM-UNet [21]	27.427	4.112	82.55	90.44	arXiv24
	LightM-UNet [32]	0.403	0.391	81.49	89.80	ICIC24
	LB-UNet [29]	0.038	0.098	82.40	90.35	MICCAI24
	ULVM-UNet [27]	0.049	0.060	83.05	90.74	arXiv24
	AULUNet-v1 (Ours)	**0.029**	**0.056**	83.65 ± 0.17	91.10 ± 0.10	-
	AULUNet-v2 (Ours)	0.031	**0.056**	84.36 ± 0.37	91.52 ± 0.22	-
	AULUNet-v3 (Ours)	**0.029**	0.069	**85.32 ± 0.25**	**92.08 ± 0.15**	-
ISIC2018	U-Net [20]	7.773	13.758	74.64	85.48	MICCAI15
	SCR-Net [26]	0.801	1.567	79.27	88.44	AAAI21
	ASwin U-Net [1]	46.910	14.181	74.62	85.46	ISBI23
	C^2SDG [11]	22.001	7.972	80.00	88.88	MICCAI23
	UNeXt-S [24]	0.302	0.103	80.29	89.07	MICCAI22
	MALUNet [22]	0.175	0.083	81.03	89.52	BIBM22
	EGE-UNet [23]	0.053	0.072	79.82	88.78	MICCAI23
	VM-UNet [21]	27.427	4.112	80.96	89.48	arXiv24
	LightM-UNet [32]	0.403	0.391	80.16	88.99	ICIC24
	LB-UNet [29]	0.038	0.098	81.22	89.64	MICCAI24
	ULVM-UNet [27]	0.049	0.060	80.64	89.29	arXiv24
	AULUNet-v1 (Ours)	**0.029**	**0.056**	81.57 ± 0.16	89.85 ± 0.10	-
	AULUNet-v2 (Ours)	0.031	**0.056**	81.90 ± 0.30	90.05 ± 0.18	-
	AULUNet-v3 (Ours)	**0.029**	0.069	**82.44 ± 0.26**	**90.37 ± 0.15**	-

Table 1: Comparative experimental results on the ISIC2017 and ISIC2018 datasets. We reproduce the results of SOTA methods with our train-val-test splits using their publicly available implementation to make a fair comparison. (↑) indicates higher is better, while (↓) indicates lower is better. Bold values indicate the best results. Our results are averaged over five independent runs.

with 23.7% fewer parameters and 29.6% less compute. Against ULVM-UNet, AULUNet-v3 achieves 2.27% higher mIoU and 1.34% higher DSC while using 40.8% fewer parameters.

We also observe a similar trend on ISIC2018. AULUNet-v3 consistently achieves the best performance, with 82.44% mIoU and 90.37% DSC. These results improve upon LB-UNet by 1.22% in mIoU and 0.73% in DSC, with a substantial reduction in complexity. Compared to ULVM-UNet, AULUNet-v3 achieves 1.80% higher mIoU and 1.08% higher DSC while maintaining the same 40.8% reduction in parameters. Even the smallest variant, AULUNet-v1, which uses only 0.029 million parameters and 0.056 GFLOPs, surpasses all baselines except AULUNet-v2 and AULUNet-v3, achieving 81.57% mIoU and 89.85% DSC.

Compared to the standard U-Net, AULUNet-v3 delivers 7.80% higher mIoU and 4.89% higher DSC on ISIC2018, while reducing the number of parameters by 99.6% (from 7.773 million to 0.029 million) and compute by 99.5% (from 13.758 to 0.069 GFLOPs). This significant reduction in complexity, combined with improved accuracy, demonstrates the capability of the proposed AULUNet for deployment in resource-constrained clinical environments and mobile platforms. All three variants offer strong trade-offs suited to varying resource budgets. A qualitative comparison in Figure 3 further illustrates the overall segmentation quality achieved by our proposed method on both datasets.

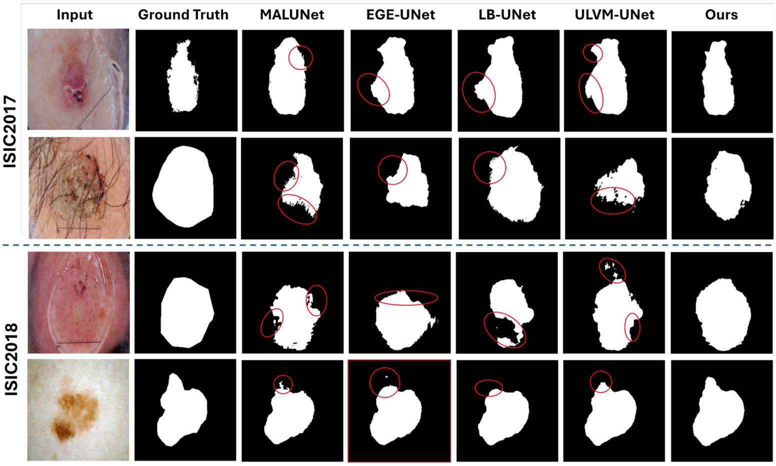

Figure 3: Qualitative comparison of AULUNet-v3 with SOTA lightweight models on ISIC2017 and ISIC2018 datasets. Red boundaries indicate segmentation errors in other methods.

Model	AKF	RGF	LSG	Params (M)↓	GFLOPs↓
Baseline				**0.008**	**0.025**
M1	✓			0.020	0.056
M2		✓		0.011	**0.025**
M3			✓	0.012	0.038
M4	✓	✓		0.024	0.056
M5	✓		✓	0.025	0.069
M6		✓	✓	0.016	0.038
AULUNet-v1	✓	✓	✓	0.029	0.056
AULUNet-v2	✓	✓	✓	0.031	0.056
AULUNet-v3	✓	✓	✓	0.029	0.069

Table 2: Ablation study on different combinations of proposed mechanisms (AKF, RGF, LSG) and their impact on model complexity (Params and GFLOPs).

3.4 Ablation Study

We conduct a comprehensive ablation study on the ISIC2018 dataset to evaluate the contribution of each component in AULUNet. Our analysis focuses on three key aspects: the effectiveness of individual modules, the impact of combined configurations, and alternative dilation strategies within the Adaptive Kernel Fusion (AKF) module. Results are summarized in Tables 2, 3, and 4. All model variants are trained and evaluated under consistent training and evaluation settings for fair comparison.

We begin with a minimal U-Net baseline composed of five encoder-decoder stages and a bottleneck with channel widths $[5, 10, 15, 25, 35, 60]$. Although highly efficient (0.008M parameters, 0.025 GFLOPs), this baseline achieves only 72.94% mIoU and 84.35% DSC,

Model	mIoU↑	DSC↑	Accuracy↑	Specificity↑	Sensitivity↑
Baseline	72.94	84.35	93.66	97.18	80.60
M1 (AKF)	80.61	89.27	95.54	97.71	87.47
M2 (RGF)	77.96	87.61	94.85	97.26	85.90
M3 (LSG)	73.69	84.85	93.87	97.32	81.04
M4 (AKF + RGF)	80.67	89.30	95.49	97.26	88.90
M5 (AKF + LSG)	81.49	89.80	95.76	97.86	87.99
M6 (RGF + LSG)	78.50	87.95	95.06	97.73	85.12
AULUNet-v1 (All)	81.57 ± 0.16	89.85 ± 0.10	95.83 ± 0.05	**98.18 ± 0.17**	87.11 ± 0.42
AULUNet-v2 (All)	81.90 ± 0.30	90.05 ± 0.18	95.82 ± 0.05	98.04 ± 0.26	87.55 ± 0.75
AULUNet-v3 (All)	**82.44 ± 0.26**	**90.37 ± 0.15**	**95.99 ± 0.04**	97.88 ± 0.15	**88.95 ± 0.72**

Table 3: Ablation results on ISIC2018 using mIoU, DSC, Accuracy, Specificity, and Sensitivity (%). Best values are bolded.

Two-Branch Fusion: (Params (M): **0.029** and GFLOPs: **0.069**)					
Variant	**mIoU↑**	**DSC↑**	**Variant**	**mIoU↑**	**DSC↑**
d = 1, d = 2 (Ours)	**82.44**	**90.37**	d = 1, d = 5	81.84	90.02
d = 1, d = 3	81.08	89.55	d = 1, d = 7	81.32	89.69
d = 2, d = 4	81.66	89.90	d = 3, d = 5	81.54	89.83
Three-Branch Fusion: Softmax-Based Adaptive Gating					
Variant	**mIoU↑**	**DSC↑**	**Params (M)↓**	**GFLOPs↓**	
d = 1, d = 2, d = 4	81.73	89.95	**0.026**	0.**079**	
d = 1, d = 3, d = 5	81.01	89.51	0.026	0.079	
d = 3, d = 5, d = 7	81.55	89.83	0.026	0.079	

Table 4: Ablation study of different dilation settings and gating strategies in AKF.

highlighting its limited capacity for spatial representation. Adding the AKF block alone (M1) significantly improves performance to 80.61% mIoU and 89.27% DSC, confirming its ability to enhance feature extraction by combining standard convolution $(3\times3, d=1)$ and dilated convolution $(3\times3, d=2)$ branches with adaptive per-channel weighting. The RGF module (M2), which injects global context through residual modulation, also improves accuracy to 77.96% mIoU and 87.61% DSC with minimal computational overhead. LSG (M3) provides limited improvement alone (73.69% mIoU), but plays an important role when used alongside AKF or RGF. For example, combining AKF and LSG (M5) achieves 81.49% mIoU and 89.80% DSC, suggesting that adaptive skip modulation strengthens the decoder's ability to refine spatial detail. Our full variants—AULUNet-v1, v2, and v3—combine all modules progressively, with AULUNet-v3 achieving 82.44% mIoU and 90.37% DSC using just 0.029M parameters and 0.069 GFLOPs.

To further validate the design of AKF, we examine multiple dilation configurations and fusion strategies. As shown in Table 4, the two-branch setup using $(d=1,2)$ offers the best trade-off between receptive field expansion and fine-grained localization, yielding the highest accuracy. Increasing the dilation to larger values $(d=3,4,5,7)$ degrades performance due to over-smoothing and diluted structural detail. We also examine several three-branch variants using softmax-based adaptive gating, which slightly reduces parameters to 0.026M but increases GFLOPs to 0.079. This configuration fails to outperform the two-branch setup, indicating that the added complexity does not improve segmentation. Therefore, these findings support our design choice: a two-branch fusion with (3×3) convolutions at $d=1\&2$ and sigmoid-based gating as a balanced approach for maintaining accuracy and efficiency.

4 Conclusion

We present AULUNet, an ultra-lightweight framework for skin lesion segmentation that delivers top-tier accuracy with minimal computational overhead—ideal for resource-limited settings like mobile health devices. By integrating adaptive multi-scale feature extraction (AKF), global context refinement (RGF), and selective feature fusion (LSG), AULUNet achieves SOTA performance with just 0.029M parameters and 0.069 GFLOPs, excelling on ISIC2017 (85.32% mIoU, 92.08% DSC) and ISIC2018 (82.44% mIoU, 90.37% DSC). Future efforts will adapt AULUNet to tasks like retinal or brain imaging and enhance its robustness against varying lighting and noise, paving the way for practical clinical deployment.

5 Acknowledgement

Thank you to the Sketch Recognition Lab (SRL) members for their continued support. This work was supported in part by NSF award 1948660.

References

[1] Ehsan Khodapanah Aghdam, Reza Azad, Maral Zarvani, and Dorit Merhof. Attention swin u-net: Cross-contextual attention mechanism for skin lesion segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

[2] Reza Azad, Moein Heidary, Kadir Yilmaz, Michael Hüttemann, Sanaz Karimijafarbigloo, Yuli Wu, Anke Schmeink, and Dorit Merhof. Loss functions in the era of semantic segmentation: A survey and outlook. *arXiv preprint arXiv:2312.05391*, 2023.

[3] Sijing Cai, Yunxian Tian, Harvey Lui, Haishan Zeng, Yi Wu, and Guannan Chen. Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quantitative imaging in medicine and surgery*, 10 (6):1275, 2020.

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[7] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by

the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[8] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[9] Md Kamrul Hasan, Md Asif Ahamad, Choon Hwai Yap, and Guang Yang. A survey, review, and future trends of skin lesion segmentation and classification. *Computers in Biology and Medicine*, 155:106624, 2023.

[10] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[11] Shishuai Hu, Zehui Liao, and Yong Xia. Devil is in channels: Contrastive single domain generalization for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–23. Springer, 2023.

[12] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[13] Nabil Ibtehaz and M Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87, 2020.

[14] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE international symposium on multimedia (ISM)*, pages 225–2255. IEEE, 2019.

[15] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519, 2019.

[16] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[18] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[19] Md Maklachur Rahman, Abdullah Aman Tutul, Ankur Nath, Lamyanba Laishram, Soon Ki Jung, and Tracy Hammond. Mamba in vision: A comprehensive survey of techniques and applications. *arXiv preprint arXiv:2410.03105*, 2024.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[21] Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

[22] Jiacheng Ruan, Suncheng Xiang, Mingye Xie, Ting Liu, and Yuzhuo Fu. Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1150–1156. IEEE, 2022.

[23] Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 481–490. Springer, 2023.

[24] Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: Mlp-based rapid medical image segmentation network. In *International conference on medical image computing and computer-assisted intervention*, pages 23–33. Springer, 2022.

[25] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.

[26] Huisi Wu, Jiafu Zhong, Wei Wang, Zhenkun Wen, and Jing Qin. Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2916–2924, 2021.

[27] Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *arXiv preprint arXiv:2403.20035*, 2024.

[28] Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li. Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th international conference on information technology in medicine and education (ITME)*, pages 327–331. IEEE, 2018.

[29] Jiahao Xu and Lyuyang Tong. Lb-unet: A lightweight boundary-assisted unet for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 361–371. Springer, 2024.

[30] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. 2015.

[31] Chunyu Yuan, Dongfang Zhao, and Sos S Agaian. Mucm-net: A mamba powered ucm-net for skin lesion segmentation. *arXiv preprint arXiv:2405.15925*, 2024.

[32] Yue Zhang, Chao Xu, Zhifan Zhang, and Jianjun Wang. Light-unet: An efficient segmentation network for medical image. In De-Shuang Huang, Zhanjun Si, and Jiayang Guo, editors, *Advanced Intelligent Computing Technology and Applications*, pages 302–313, Singapore, 2024. Springer Nature Singapore. ISBN 978-981-97-5597-4.

[33] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2021.

[34] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.