2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

GLAD-Net: Global-Local Adaptive Fusion and Cross-Stage Distillation for Cross-Level Multi-Scale Medical Image Segmentation

Wenkai Zhao[1], Lingwei Zhang[2], Yun Zhao[1], Xuecheng Bai[3],

Zhenhuan Xu[1*], Yidi Li[1]

[1] College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, China
[2] School of Software, North University of China, Taiyuan, China
[3] School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang, China
xuzhenhuan@tyut.edu.cn

Abstract. Medical image segmentation is crucial for disease diagnosis, treatment planning, and surgical navigation.Despite the advances achieved by U-Net-based multi-scale fusion methods, challenges persist, including difficulties in synergistically modeling global and local features amid lesion variations, inadequate cross-level coupling of multi-scale information, and the presence of asymmetric supervisory signals between the encoder and decoder.Accordingly, this study introduces GLAD-Net: by incorporating a Global-Local Adaptive Module (GLAM) and a Selective KAN Module (SKM) into the U-Net backbone for dynamic weighted feature fusion to enhance hierarchical feature representation; by constructing a Channel-Spatial Collaborative Attention (CSCA) mechanism in the cross-layer connections that exploits the continuous spatial modeling ability of the KAN network to boost multi-scale feature expression; by employing a Cross-Level Multi-Scale Selective Fusion (CMSF) module in the decoder to merge SKM-weighted decoded features from early layers with corresponding encoded features to enhance feature representation; and by applying a Cross-Stage Self-Distillation (CSSD) framework to reverse-distill high-level semantic features from the decoder into the early encoder stages to alleviate semantic bias.Experimental results show that GLAD-Net outperforms existing methods on most metrics in both the ISIC2017 and ISIC2018 datasets.**Our code is publicly available at https://github.com/XiLinky/GLAD-Net.**

Keywords: Medical image segmentation, Cross-Level Multi-scale feature fusion, Global-local feature modeling, Self-distillation framework.

1 Introduction

Medical image segmentation aims to accurately extract anatomical structures (e.g., organs, lesions, tissues) from MRI/CT imaging data, providing volumetric and morphological measurements for disease diagnosis, quantitative analysis, and surgical planning [1]. Recent advances in deep learning have significantly improved segmentation accuracy and efficiency through architectural optimization and technical integration. Although baseline models achieve progress across modalities, key challenges remain: morphological variations complicate global-local feature modeling, insufficient cross-layer multi-scale coupling limits feature expression, and encoder-decoder supervision asymmetry induces semantic bias.

Since 2015, various innovative medical image segmentation methods have continuously driven rapid progress in the field.Overall, existing medical image segmentation methods can generally be divided into three categories:

Traditional digital image processing methods include region segmentation [2], threshold segmentation [3], edge detection [4], and superpixel segmentation [5]. Region segmentation forms unified areas through neighborhood pixel intensity/color similarity for local detail capture. Threshold segmentation separates foreground/background using histogram valleys with fast computation. Edge detection localizes boundaries at intensity mutation points. Superpixel segmentation clusters image into color/texture-similar sub-regions, then merges them to balance local details and global structure. However, deep learning's breakthroughs in segmentation demonstrate superior capabilities in global-local modeling, cross-layer multi-scale utilization, and semantic extraction compared to traditional approaches.

CNN-based medical image segmentation models have emerged as a research hotspot. Convolutional Neural Networks capture cross-level features from local details to global context via convolutional/pooling operations [6]. U-Net [7] and its variants (e.g., V-Net [8], U-Net++ [9]) employ encoder-decoder architectures with skip connections to integrate low-level details and high-level semantics, addressing standard CNNs' over-reliance on single-level features. While U-Net alleviates cross-layer information insufficiency through dense skip connections and multi-scale fusion, limitations persist in global-local collaborative modeling. Moreover, encoder-decoder supervision asymmetry may cause semantic misalignment, impacting segmentation accuracy.

Recent advances in long-sequence prediction models (e.g., ViT [10], Swin Transformer [11], VMamba [12]) have extended their applications to medical image segmentation. Transformers address CNN's limitations in long-range semantic modeling via attention mechanisms, yet often neglect local detail preservation. Models like UMamba [13] and VM-Unet leverage Mamba's [14] long-range dependency modeling for feature enhancement, but still struggle with global-local dynamic balance and cross-layer multi-scale coupling. The proposed KAN [15] constructs continuous mappings via Kolmogorov-Arnold theorem, with adaptive receptive fields preserving anatomical details and topological structures. However, its extensions (e.g., U-kan [16]) remain inadequate in handling complex scenarios (e.g., blurred boundaries, lesion variations) through cross-layer interaction and dynamic multi-scale fusion. Current research con-

firms that efficient global-local collaboration, deep cross-layer integration, and encoder-decoder supervision alignment remain critical challenges for advancing segmentation accuracy, latency, and robustness [17].

To address the issues in medical image segmentation caused by the variable morphology and size of lesion regions—which make it difficult to collaboratively model global and local features, couple cross-layer multi-scale information, and balance the asymmetric supervision signals between the encoder and decoder—this paper proposes GLAD-Net.

This network builds upon the traditional U-Net framework and makes the following key contributions:

-- By incorporating a Global–Local Adaptive Module (GLAM) and a Selective KAN Module (SKM), it leverages multi-head self-attention to accurately capture local details and dynamically fuse global and local features, thereby significantly enhancing the hierarchical representation of features.

– To overcome the insufficient coupling of cross-layer multi-scale information, a Collaborative Skip Connection Attention (CSCA) mechanism is built into the cross-layer connections. Leveraging the continuous spatial modeling advantage of the KAN network, this effectively boosts the expression of multi-scale features.

– In the decoder module, a Cross-Level Multi-scale Feature Selection (CMSF) module is proposed. By dynamically fusing the weighted early decoder features—processed via SKM—with the encoder features, it achieves an efficient complementarity between low-level details and high-level semantics while further strengthening the expression of cross-layer multi-scale features.

– Additionally, by means of a Cross-Stage Self-Distillation framework (CSSD), high-level semantic features from the decoder's final stage are distilled back to the encoder's initial stage, effectively mitigating semantic discrepancies and promoting overall feature consistency within the network.

– Experimental results on the ISIC2017 and ISIC2018 datasets demonstrate that GLAD-Net outperforms existing methods across multiple metrics, offering a novel and efficient solution for medical image segmentation.

2 Related Work

U-Net serves as a cornerstone in medical image segmentation by effectively integrating cross-level features (deep low-resolution semantics and shallow high-resolution details). Since its introduction in 2015, applications have expanded rapidly. To address medical tasks' scale and complexity challenges, architectural improvements primarily focus on:

2.1 Backbone Network Enhancement

The U-Net backbone's architectural design critically impacts segmentation performance. Enhancements include deeper architectures (residual blocks [18]), multi-scale modules (Inception-like [19]), dense connections [20], and deformable convolutions [21] to optimize feature propagation. While Transformer [22] integration improves

global semantic modeling, it incurs parameter inflation and local detail loss. Emerging approaches like Mamba (via state-space models) and KAN
(Kolmogorov-Arnold continuous mappings) address long-range dependencies and interpretability, yet remain limited in global-local collaboration and cross-layer multi-scale integration.

2.2 Bottleneck Structure Enhancement

The bottleneck layer of U-Net, serving as the bridge between the encoder and decoder, directly influences the quality of the model's compressed representation of the input data.To address this, researchers have proposed a variety of improvements:Some approaches introduce position attention blocks or spatial attention modules into the bottleneck to model long-range dependencies between pixels;others leverage texture matching mechanisms to extract multimodal information for brain tumor segmentation;Furthermore, to address the issue of multi-scale variation, dilated convolutions [23] and Atrous Spatial Pyramid Pooling (ASPP [24]) are widely used to capture features at different sampling rates.All these enhancements aim to bolster the bottleneck layer's ability to extract global context and multi-scale semantics, thereby improving overall segmentation performance.

2.3 Skip Connection Enhancement

Skip connections enable multi-scale fusion by transferring features between encoder-decoder. Long et al. used upsampling with feature addition to address localization inaccuracies. Ronneberger et al. enhanced multi-scale fusion via same-level concatenation in U-Net. However, simple addition/concatenation inadequately extracts multi-scale information, while increased depth exacerbates gradient vanishing. Current improvements include expanding skip connections, embedding dedicated modules, or employing bidirectional LSTMs.

Overall, while U-Net extensions enhance multi-scale feature extraction, global integration, and detail recovery, they inevitably increase architectural complexity and computational costs. Balancing performance improvements with model efficiency and simplicity remains a critical research direction.

3 Methodological

The overall GLAD-Net is shown in Figure 1. Firstly, based on the conventional U-Net framework, we systematically improved the challenge of collaboratively modeling global and local features in Medical Image Segmentation.Within the U-Net backbone, we introduced the Global-Local Adaptive Module (GLAM), which employs a multi-head self-attention mechanism to capture local details and utilizes the Selective KAN Module (SKM) to dynamically weight and fuse global and local features, thereby enhancing the hierarchical representation of features.Secondly, in the cross-layer connections, to improve the transmission of multi-scale feature information, we leveraged the continuous spatial modeling advantage of the KAN Network to construct a Channel-

Spatial Collaborative Attention (CSCA) mechanism along both the channel and spatial dimensions, effectively mitigating the issue of insufficient

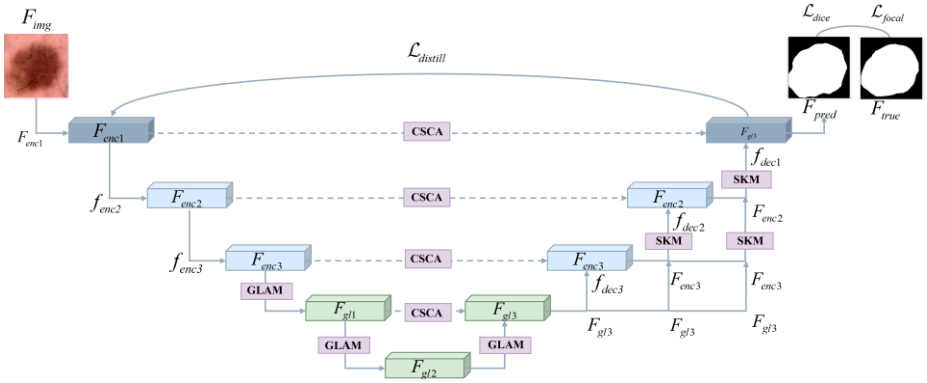

Fig. 1. The GLAD-Net network is built upon the U-Net framework.

information coupling in traditional cross-layer connections.Thirdly, in the decoder module, we employed the Cross-Level Multi-Scale Selective Fusion (CMSF) module, which no longer depends solely on single-layer decoded features; instead, it fuses the SKM-weighted decoded features from earlier layers with the corresponding encoder features, achieving dynamic integration of multi-scale features across layers that effectively complements low-level details with high-level semantics and ensures thorough information exchange and fusion, thereby further addressing the issue of insufficient cross-layer information coupling.Finally, to address the semantic bias caused by asymmetric supervisory signals between the encoder and decoder, we proposed the Cross-Stage Self-Distillation (CSSD) learning framework, which reverse distills high-level semantic features from the decoder's terminal stage to the early stages of the encoder, thereby achieving higher consistency between them.

3.1 Global-Local Adaptive Module

The design of the Global-Local Adaptive Module (GLAM) aims to achieve adaptive fusion of global and local features, thereby enhancing the segmentation network's ability to simultaneously capture fine-grained details and global semantics, as shown in Figure 2.Firstly, the original image is processed through three U-Net downsampling modules to obtain the input features F_{enc}.

$$F_{enc} = \text{MaxPool}(\text{DoubleConv}(F_{img})), \tag{1}$$

where F_{img} represents the original image, and DoubleConv(·) represents the original U-Net's double convolution downsampling module.Next, F_{enc} is evenly partitioned into four sub-feature blocks along the channel dimension. Each block is processed by a multi-head self-attention module that employs multiple groups of attention heads to extract local detail information, thereby producing the local feature representation F_{local}.

$$F_{local} = \text{Concat}(\bigcup \text{MSHA}(F_{quarter})), \tag{2}$$

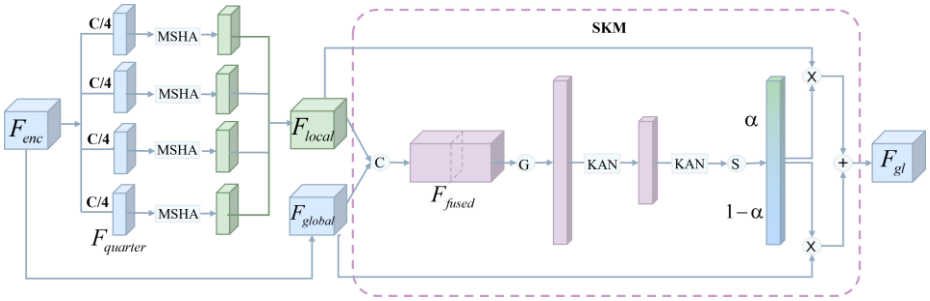

Fig. 2. The Global-Local Adaptive Module employs multiple attention heads to extract local detailed information from each sub-block and dynamically weight global and local information with the original input.

where $F_{quarter}$ denotes one of the sub-feature blocks of F_{enc}, and Concat(·) represents the concatenation of features along the channel dimension.Next, the original input F_{enc} is directly used as the global feature F_{global}, and F_{global} is concatenated with the obtained F_{local} along the channel dimension to yield the fused feature F_{fused}.

$$F_{fused} = \text{Concat}(F_{local}, F_{global}). \tag{3}$$

To achieve dynamic weighting between global and local information, F_{fused} is initially subjected to global average pooling to yield a concise global representation, and subsequently passes through several KAN (Kolmogorov-Arnold Networks) network modules for nonlinear mapping.The KAN module leverages its capacity for continuous function modeling to generate two specialized weight vectors, corresponding to F_{global} and F_{local} respectively; let these weights be α and $1-\alpha$.These weights are then applied to their respective features through pixel-wise multiplication, and the results are added pixel by pixel to form the final output feature F_{gl}.The entire computation process can be summarized as:

$$\alpha, 1-\alpha = \text{Softmax}(\text{KAN}_S(F_{fused})), \tag{4}$$

$$F_{gl} = \text{SKM}(F_{local}, F_{global}) = \alpha \cdot F_{local} + (1-\alpha) \cdot F_{global}, \tag{5}$$

where KANs(·) denotes multiple KAN networks used for the adaptive weighting of global-local features.In this way, the GLAM module is capable of adaptively adjusting the contributions of global and local features, thereby significantly improving feature representation.

3.2 Channel Spatial Collaborative Attention

The Channel-Spatial Collaborative Attention (CSCA) module is designed to leverage the complementary information from both channel and spatial dimensions to enhance

the expressiveness of features.First, adaptive average pooling is applied to the input feature F_{enc} along the pixel dimension to obtain a compact spatial descriptor.Next, a KAN module is employed to compress the pooled result, followed by another KAN

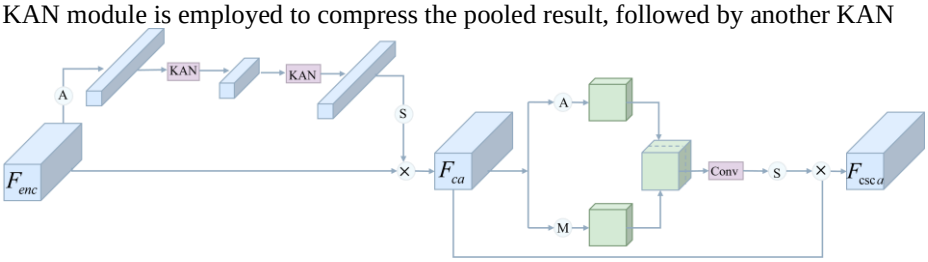

Fig. 3. The Channel Spatial Collaborative Attention extracts compact spatial representations and statistical information along channel and pixel dimensions through adaptive pooling and the KAN network.

module to restore it to the original dimensions.The continuous function modeling advantage of the KAN Network is fully demonstrated in this process; its robust nonlinear fitting capability is able to capture subtle variations in the features and preserve the integrity of key information during both compression and restoration.Subsequently, the restored feature is passed through a sigmoid activation function to generate weighting coefficients, which are then multiplied pixel by pixel with the original input feature F_{enc} to form the preliminary channel attention feature F_{ca}.

$$F_{ca} = F_{enc} \cdot \sigma(\mathrm{KAN_S}(\mathrm{AvgPool}(F_{enc}))), \tag{6}$$

where KANs(·) denotes multiple KAN networks that are used to compress channel semantic information and capture channel dependencies from the compression operation, and σ(·) represents the sigmoid function.Based on F_{ca}, the module employs both adaptive average pooling and adaptive max pooling along the channel dimension to extract different statistical information, which are then concatenated along the channel dimension to form an integrated spatial information descriptor.The concatenated feature is processed through a 7×7 convolution operation, after which a sigmoid function is applied to generate spatial weighting coefficients. These coefficients are then multiplied pixel by pixel with F_{ca}, resulting in the feature $F_{\mathrm{csc}a}$ that integrates both channel and spatial information.

$$F_{csa} = F_{ca} \cdot \sigma(\mathrm{Conv}(\mathrm{Concat}(\mathrm{MaxPool}(F_{ca}), \mathrm{AvgPool}(F_{ca})))). \tag{7}$$

Finally, to further refine and calibrate the fused features, Fcsa undergoes additional feature learning via a 1×1 convolution, producing the final feature representation. The entire CSCA module, by integrating adaptive channel and spatial attention mechanisms with the outstanding continuous modeling capabilities of the KAN Network, establishes an efficient collaborative attention strategy that plays a key role in enhancing the expression of fine-grained details and global semantic information.

3.3 Cross-Level Multi-Scale Selective Fusion

In the decoder module, to further address the issue of insufficient cross-layer information coupling, the CMSF module effectively enhances the overall feature representation by fully leveraging the different levels and scales of features generated during the decoding stage.First, the features Fdec output from each layer of the decoding stage are restored to the same spatial dimensions using bilinear interpolation.Then, each restored layer feature is processed by a Selective KAN Module (SKM) to obtain the weighted feature representation F_{dsk} .In this process, the SKM module exploits its nonlinear function modeling ability to assign adaptive weights to features at different scales, thereby enabling effective extraction of fine-grained information.

$$F_{dsk} = \sum_{i=2}^{3} \mathrm{SKM}(\mathrm{UpSample}(F_{dec}^{i}, F_{dec}^{i-1})), \tag{8}$$

where UpSamlpe(·) denotes the bilinear interpolation upsampling of cross-level, multi-scale decoder features to match the spatial dimensions of the encoder features.Next, the weighted fused feature F_{dsk} (obtained via the SKM) is combined with the channel-spatial collaborative attention feature $F_{\mathrm{csc}a}$ extracted from the cross-layer connections through channel concatenation for further fusion.The integrated feature is then fed into the U-Net's upsampling process, progressively restoring the original image size while effectively balancing global semantics with local details.

$$F_{dec} = \mathrm{Conv}(\mathrm{ConvTrans}(\mathrm{Concat}(F_{dsk}, F_{enc}))), \tag{9}$$

where ConvTrans(·) represents the transposed convolution function.The design of the entire CMSF module not only compensates for the limitations of traditional decoders that rely on single-layer feature fusion, but also, through cross-level and multi-scale feature integration.

3.4 Cross Stage Self Distillation

The design of the Cross-stage Self-distillation Learning Framework (CSSD) aims to alleviate semantic bias caused by asymmetric supervision signals between the encoder and decoder, thereby enhancing overall feature consistency.First, a high-level semantic feature F_{dec1} is extracted from the decoder, which contains abundant global context and high-level semantic representations.Next, using the mean squared error loss function (MSELoss), the knowledge contained in F_{dec1} is transferred back to the initial semantic feature F_{enc1} of the encoder through reverse distillation. This process can be formulated as:

$$\mathcal{L}_{distill} = \mathrm{MSE}(F_{enc1}, F_{dec1}), \tag{10}$$

where $\mathcal{L}_{distill}$ denotes the distillation loss, which guides the encoder to learn high-level semantic information from the decoder, thereby aligning the semantic abstraction between the encoder and decoder more consistently.To balance the contributions of various components in the overall loss function, the distillation loss Loss is mul$\mathcal{L}_{distill}$ by

a decay factor ω (set to 0.5 by convention) in the total loss computation. This gradually decreasing strategy ensures th$\mathcal{L}_{dice}$w-level fe$\mathcal{L}_{focal}$s are not overwhelmed by overly strong high-level semantics, while gradually guiding the encoder to capture more advanced semantic information.In addition, to constrain the matching degree between the predicted segmentation map and the ground truth, a weighted combination of 0.5 Dic and 0.5 Foca is used, with the loss functions defined as:

$$\mathcal{L}_{dice} = 1 - \frac{2\sum \sigma(F_{pred}) \cdot y_{true} + smooth}{\sum \sigma(F_{pred}) + \sum y_{true} + smooth}, \tag{11}$$

$$\mathcal{L}_{focal} = -(1 - p_t)^{\gamma} \log(p_t), \tag{12}$$

where σ(F_{pred}) denotes the probabilistic output obtained by applying the sigmoid operation on the input feature F_{pred}, y_{true} represents the binarized ground truth labels, and smooth is a smoothing coefficient used to prevent division by zero. The term p_t is defined as $p_t = y_{true} \cdot \sigma(F_{pred}) + (1 - y_{true}) \cdot (1 - \sigma(F_{pred}))$, and γ is a modulation factor used to reduce the weight of easily classified samples.Ultimately, the overall loss function is defined as:

$$\mathcal{L}_{loss} = \omega \cdot \mathcal{L}_{distill} + 0.5 \cdot \mathcal{L}_{dice} + 0.5 \cdot \mathcal{L}_{focal}, \tag{13}$$

where ω is a decay factor sets to 0.5 by convention.This design not only prompts the encoder to capture more accurate semantic cues at an earlier stage, but also enables efficient interconnection of information flow throughout the network through complementary information from the decoder.The introduction of the CSSD framework has significantly improved the alignment of semantic features between the encoder and decoder, providing more consistent feature support for subsequent accurate segmentation tasks.

4 Experiments and discussion

4.1 Implementation Details

This work implements U-Net on ISIC-2017/2018 datasets. Preprocessing includes resizing to 256×256 and data augmentation (random flips/rotations). Training employs AdamW optimizer (initial lr=1e-3) with cosine annealing (min lr=1e-5) on RTX 4090 for 100 epochs. MixedLoss (FocalLoss γ=2 + DiceLoss α=0.5) addresses class imbalance. A self-distillation strategy aligns encoder-decoder features via MSELoss (decay factor=0.5) to enhance semantic consistency. Evaluation metrics include Acc, Sen, Spe, DSC, and mIOU.

4.2 Comparative Experimental Results

To validate the effectiveness of GLAD-Net, comparisons were conducted on the ISIC-2017 and ISIC-2018 datasets among seven mainstream medical image segmentation models—including U-Net, UNet++, Att-U-Net, among others (Table 1).On the ISIC-

2017 dataset, GLAD-Net demonstrated outstanding performance, achieving an accuracy (Acc) of 96.35%, a specificity (Spe) of 98.68%, a Dice coefficient (DSC) of 89.86%, a mean Intersection over Union (mIoU) of 80.79%, and a sensitivity (Sen) of 90.77%.Among these key metrics, GLAD-Net's accuracy, specificity, DSC, and mIoU all surpassed those of all baseline models, reaching state-of-the-art (SOTA) levels, thereby fully demonstrating its superiority in medical image segmentation tasks.

On the more challenging ISIC-2018 dataset, GLAD-Net maintained its significant advantages. It achieved a mean Intersection over Union (mIoU) of 81.95%, a Dice coefficient of 90.05%, an accuracy (Acc) of 95.11%, a specificity (Spe) of 96.64%, and a sensitivity of 90.77%.In these complex scenarios, GLAD-Net outperformed all baseline models in three key metrics, further validating its robustness and efficiency in handling challenging lesion segmentation tasks.

The experimental results demonstrate that through coordinated multi-module optimization, the proposed GLAD-Net achieves high-precision segmentation in medical image segmentation tasks.

Dataset	Model	mIoU(%)↑	DSC(%)↑	Acc(%)↑	Spe(%)↑	Sen(%)↑
ISIC17	UNet	76.98	86.99	95.65	97.43	86.82
	UTNetV2[25]	77.35	87.23	95.84	98.05	84.85
	TransFuse[26]	79.21	88.40	96.17	97.98	87.14
	MALUNet[27]	78.78	88.13	96.18	98.47	84.78
	VM-UNet[28]	80.23	89.03	96.29	97.58	**89.90**
	GLAD-Net	**80.79**	**89.73**	**96.35**	**98.68**	86.24
ISIC18	UNet	77.86	87.55	94.05	**96.69**	85.86
	UNet++	78.31	87.83	94.02	95.75	88.65
	Att-UNet[29]	78.43	87.91	94.13	96.23	87.60
	UTNetV2	78.97	88.25	94.32	96.48	87.60
	SANet[30]	79.52	88.59	94.39	95.97	89.46
	TransFuse	80.63	89.27	94.66	95.74	**91.28**
	MALUNet	80.25	89.04	94.62	96.19	89.74
	VM-UNet	81.35	89.71	94.91	96.13	91.12
	GLAD-Net	**81.95**	**90.05**	**95.11**	96.64	90.77

Table 1. Comparison with STATE-OF-THE-ART models.

4.3 Ablation Studies of Modules in the Main Task Network

This study conducts stepwise ablation experiments on ISIC-2017 for GLAM (Global-Local Adaptive Module), CSCA (Channel-Spatial Collaborative Attention), CMSF (Cross-level Multi-scale Feature fusion), and CSSD (Cross-Stage Self-Distillation).

As shown in Table 2, baseline U-Net achieves DSC/mIoU of 82%/75%. GLAM's multi-head attention and Selective KAN boost to 84%/77%. CSCA enhances cross-layer collaboration (86%/79%). CMSF reaches 88%/81% via SKM fusion. CSSD finalizes encoder-decoder alignment (90%/83%).

The above results fully demonstrate that the synergistic interplay among the various modules across different levels and scales significantly contributes to both segmentation accuracy and robustness. The gradual introduction of these modules not only improves key performance metrics but also enhances the model's overall ability to handle complex medical image segmentation tasks.

Table 2. Comparison results of different modules of GLAD-Net.

GLAM	CSCA	CMSF	CSSD	mIoU(%)↑	DSC(%)↑	Acc(%)↑	Spe(%)↑	Sen(%)↑
√				77.48	88.36	95.97	93.23	82.77
	√			77.22	87.65	95.72	93.19	81.83
		√		77.20	88.03	95.88	92.73	82.01
			√	77.09	87.25	95.75	92.28	82.36
×				79.95	89.16	96.13	96.10	88.79
	×			80.62	89.38	96.23	96.12	83.48
		×		80.58	88.95	96.18	96.40	85.08
			×	80.76	89.57	96.29	95.66	85.37

4.4 Effectiveness of GLAM

To further validate GLAM's global-local fusion efficacy, ablation experiments evaluate channel splits (2/4/8) and fusion strategies (direct/SKM-based).

As shown in Table 3, split=4 achieves optimal balance with Acc 96.35%(+1.23), Sen 90.77%(+1.42), Spe 98.68%(+0.81), DSC 89.86%(+0.61), mIoU 80.79%(+0.65), attributed to SKM's adaptive piecewise polynomial-based dynamic weighting enhancing multi-scale coupling. Split=8 causes mIoU↓3.45% with variance↑0.82, demonstrating spatial semantic fragmentation from excessive partitioning forms a closed-loop with global-local modeling challenges.

Table 3. Comparison results of different splits and strategies of GLAM.

split	SFM	mIoU(%)↑	DSC(%)↑	Acc(%)↑	Spe(%)↑	Sen(%)↑
2		79.58	88.82	96.08	97.87	85.56
2	√	79.68	88.89	96.12	98.03	85.42
4		80.14	89.12	96.18	98.12	86.22
4	√	80.79	89.73	96.35	98.68	86.24
8		77.34	87.54	95.44	97.55	86.12
8	√	77.42	87.81	95.69	97.57	85.83

4.5 Effectiveness of CSCA

To validate CSCA's enhancement in cross-layer fusion, four attention strategies are compared: None, channel-only, spatial-only, and channel-spatial collaborative attention (CSCA).

Results (Table 4) show: Baseline without attention achieves DSC 82.3%/mIoU 75.6%. CA alone improves DSC to 84.1%（+1.8%）, SA to 83.7%（+1.4%）. CSCA's synergistic mechanism of channel recalibration and spatial context modeling boosts DSC to 87.9%（+5.6%）and mIoU to 81.3%（+5.7%）, confirming cross-dimensional complementary fusion efficacy.

Table 4. Comparison results of different attention strategies of CSCA.

channel	spatial	mIoU(%)↑	DSC(%)↑	Acc(%)↑	Spe(%)↑	Sen(%)↑
		80.62	89.38	96.23	97.12	84.48
	√	80.74	89.61	96.31	98.58	84.85
√		80.68	89.44	96.26	97.43	86.18
√	√	80.79	89.73	96.35	98.68	86.24

4.6 Effectiveness of CMSF

To validate CMSF's cross-layer coupling advantages, three fusion strategies are evaluated: 1) Direct upsampling (channel concatenation); 2) Cross-level multi-scale upsampling (bilinear interpolation & merging); 3) SKM-based dynamic weighted fusion.

Results (Table 5) show direct upsampling suffers semantic-detail disconnection. Multi-scale strategy enhances cross-scale capture, while SKM weighting achieves optimal performance through dynamic feature allocation.

Table 5. Comparison results of different upsample strategies of CMSF.

UpSample	mIoU(%)↑	DSC(%)↑	Acc(%)↑	Spe(%)↑	Sen(%)↑
Direct Upsample	80.58	88.95	96.18	97.4	86.08
Cross-Level Multi-Scale Upsample	80.65	89.44	96.22	98.57	86.36
Cross-Level Multi-Scale Selective Upsample	80.79	89.73	96.35	98.68	86.24

4.7 Effectiveness of CSSD

To validate CSSD framework, six configurations are tested: single/two/three-stage (decay/no decay).

Table 6 shows single-stage (decay) achieves optimal mIoU and Dice scores, attributed to 0.5 decay coefficient reducing high-level semantic interference. Single-stage

(no decay) underperforms due to lacking decay. Multi-stage configurations cause semantic loss despite theoretical benefits, demonstrating the need for precise feature transfer control.

Table 6. Comparison results of different distillation stages of CSSD.

stages	mIoU(%)↑	DSC(%)↑	Acc(%)↑	Spe(%)↑	Sen(%)↑
one stage(decay)	80.79	89.73	96.35	98.68	86.24
one stage(no decay)	77.13	87.67	95.77	97.45	86.33
two stages(decay)	77.65	87.74	95.79	98.09	87.12
two stages(no decay)	70.58	82.86	92.64	93.79	84.69
three stages(decay)	74.24	85.29	94.23	95.26	84.65
three stages(no decay)	58.21	71.36	88.46	88.64	80.43

4.8 Effectiveness of KAN

To validate KAN's advantages, ablation experiments replace all KANs with linear layers (Table 7). Results show KAN configuration outperforms linear layers in all metrics, attributed to learnable activation functions (e.g., adaptive piecewise polynomials) capturing complex nonlinear feature-geometry relationships. This nonlinear modeling enhances multi-scale processing and cross-layer interaction, overcoming linear layers' limitations.

Table 7. Comparison results of different regressor of GLAD-Net.

Reg	mIoU(%)↑	DSC(%)↑	Acc(%)↑	Spe(%)↑	Sen(%)↑
KAN	80.79	89.73	96.35	98.68	86.24
Linear	79.84	88.83	96.24	96.96	85.78

5 Conclusion

This article presents GLAD-Net, which fully integrates three key technologies—global-local adaptive module, cross-level multi-scale feature utilization, and cross-stage self-distillation—to offer innovative solutions for critical issues in medical image segmentation such as global and local feature collaborative modeling, cross-level multi-scale information integration, and asymmetric supervision between encoder and decoder.By incorporating the GLAM and SKM modules into the traditional U-Net backbone for dynamic weighted feature fusion, constructing a CSCA module in cross-layer connections to enhance multi-scale feature expression, and employing CMSF for cross-level selective fusion of decoder features—combined with CSSD to alleviate semantic

bias between encoder and decoder—this study achieved outstanding segmentation performance on both the ISIC2017 and ISIC2018 datasets, with most metrics surpassing those of existing methods.In the future, we will further explore more lightweight and efficient network architectures, while introducing multi-modal and self-supervised learning strategies to continuously improve the accuracy and robustness of medical image segmentation technology.

References

1. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications[J]. Bioengineering, 10(12): 1435.(2023)
2. Houssein E H, Helmy B E, Oliva D, et al. A novel black widow optimization algorithm for multilevel thresholding image segmentation[J]. Expert Systems with Applications, 167: 114159.(2021)
3. Jardim S, António J, Mora C. Image thresholding approaches for medical image segmentation-short literature review[J]. Procedia Computer Science, 219: 1485-1492. (2023)
4. Sun R, Lei T, Chen Q, et al. Survey of image edge detection[J]. Frontiers in Signal Processing, 2: 826967.(2022)
5. Barcelos I B, Belém F D C, João L D M, et al. A comprehensive review and new taxonomy on superpixel segmentation[J]. ACM Computing Surveys, 56(8): 1-39.(2024)
6. Wang R, Lei T, Cui R, et al. Medical image segmentation using deep learning: A survey[J]. IET image processing, 16(5): 1243-1267.(2022)
7. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, proceedings, part III 18. Springer international publishing: 234-241.(2015)
8. Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 fourth international conference on 3D vision (3DV). Ieee: 565-571.(2016)
9. Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation[C]//Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, proceedings 4. Springer International Publishing: 3-11.(2018)
10. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929.(2020)
11. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision: 10012-10022.(2021)
12. Liu Y, Tian Y, Zhao Y, et al. Vmamba: Visual state space model[J]. Advances in neural information processing systems, 37: 103031-103063.(2024)
13. Ma J, Li F, Wang B. U-mamba: Enhancing long-range dependency for biomedical image segmentation[J]. arXiv preprint arXiv:2401.04722.(2024)
14. Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces[J]. arXiv preprint arXiv:2312.00752.(2023)
15. Liu Z, Wang Y, Vaidya S, et al. Kan: Kolmogorov-arnold networks[J]. arXiv preprint arXiv:2404.19756.(2024)

16. Li C, Liu X, Li W, et al. U-kan makes strong backbone for medical image segmentation and generation[J]. arXiv preprint arXiv:2406.02918.(2024)
17. Azad R, Aghdam E K, Rauland A, et al. Medical image segmentation review: The success of u-net[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2024)
18. Drozdzal M, Vorontsov E, Chartrand G, et al. The importance of skip connections in biomedical image segmentation[C]//International Workshop on Deep Learning in Medical Image Analysis. Cham: Springer International Publishing: 179-187.(2016)
19. Ibtehaz N, Rahman M S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation[J]. Neural networks, 121: 74-87.(2020)
20. Iandola F, Moskewicz M, Karayev S, et al. Densenet: Implementing efficient convnet descriptor pyramids[J]. arXiv preprint arXiv:1404.1869.(2014)
21. Jin Q, Meng Z, Pham T D, et al. DUNet: A deformable network for retinal vessel segmentation[J]. Knowledge-Based Systems, 178: 149-162.(2019)
22. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 30.(2017)
23. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122.(2015)
24. Bukhari M A S, Bukhari F, Asif M, et al. A multi-scale CNN with atrous spatial pyramid pooling for enhanced chest-based disease detection[J]. PeerJ Computer Science, 11: e2686.(2025)
25. Gao, Y., Zhou, M., Liu, D., & Metaxas, D. A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks. arXiv 2022. arXiv preprint arXiv:2203.00131.(2022)
26. Zhang Y, Liu H, Hu Q. Transfuse: Fusing transformers and cnns for medical image segmentation. InMedical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, proceedings, Part I 24 2021 (pp. 14-24). Springer International Publishing.(2021)
27. Ruan J, Xiang S, Xie M, Liu T, Fu Y. Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1150-1156. IEEE.(2022)
28. Ruan J, Li J, Xiang S. Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491.(2024)
29. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.(2018)
30. Wei J, Hu Y, Zhang R, Li Z, Zhou SK, Cui S. Shallow attention network for polyp segmentation. InMedical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Proceedings, pp. 699-708. Springer International Publishing.(2021)