AGTCNet: Hybrid Network Based on AGT and Curvature Information for Skin Lesion Detection

Zhiwei Dong
School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China
2022420043@sdtbu.edu.cn

Genji Yuan
School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China
yuangenji@sdtbu.edu.cn

Jinjiang Li
School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China
lijinjiang@sdtbu.edu.cn

Abstract

Early detection and diagnosis of skin cancers is essential to improve patient survival. However, traditional diagnostic methods have limitations due to the complexity and diversity of skin lesions. Although deep learning-based skin disease detection methods are available, the ambiguity of the boundaries of skin lesion regions may lead to model neglect and misclassification, generating suboptimal results and affecting clinical decisions. To address this problem, this paper proposes a hybrid network based on Adaptive Grouped Transformer (AGT) and curvature information fusion for skin lesion detection, called AGTCNet. AGTCNet enhances the network's adaptive multi-scale learning capability by introducing AGT. In addition, a curvature-based guidance enhancement module (CGEM) is proposed in this paper, which utilizes the curvature information to effectively guide the model in enhancing its capture of complex lesion edge information. To further optimize the model performance, the deep supervision mechanism is used to dynamically calculate the loss at each stage and adjust the learning strategy based on the loss feedback. Through comprehensive experimental validation on the ISIC2016, ISIC2017, and PH2 skin lesion segmentation datasets, the results show that AGTCNet significantly outperforms the existing mainstream methods on all datasets, and especially exhibits excellent performance in detailed feature processing and fuzzy region segmentation.

Keywords: ***Skin Lesion Segmentation Adaptive Grouped Transformer Curvature Information Deep Supervision***

1. Introduction

Medical image segmentation [1] aims to separate different anatomical structures and tissues from an image to help physicians accurately locate and measure lesions or abnormal regions, which plays a key role in disease diagnosis and treatment. In recent years, skin diseases are still a serious health threat, but they may mislead doctors' diagnosis due to the variability and similarity of their visual features. Therefore, clinicians need more detailed information to support their decisions.

Traditional medical image segmentation techniques help clinicians understand images more intuitively and make accurate diagnoses by extracting image features. However, these methods rely on hand-designed features, and their performance is affected by subjective factors and domain knowledge, leading to a decrease in the accuracy and consistency of segmentation results. In addition, when dealing with medical images with complex textures, traditional methods are difficult to effectively capture and abstract their complex features, exposing technical limitations.

To address the limitations, the application of deep learning methods in the medical field is rapidly expanding, especially in medical image segmentation showing higher accuracy and robustness. Deep learning methods are free from the limitation of hand-designed features and can automatically learn and extract complex textural and structural features from large amounts of data. Convolutional neural networks (CNNs)[2] perform particularly well in medical image segmentation tasks, where features are usually extracted through convolutional layers and further abstracted through fully connected layers. CNN architectures, rep-

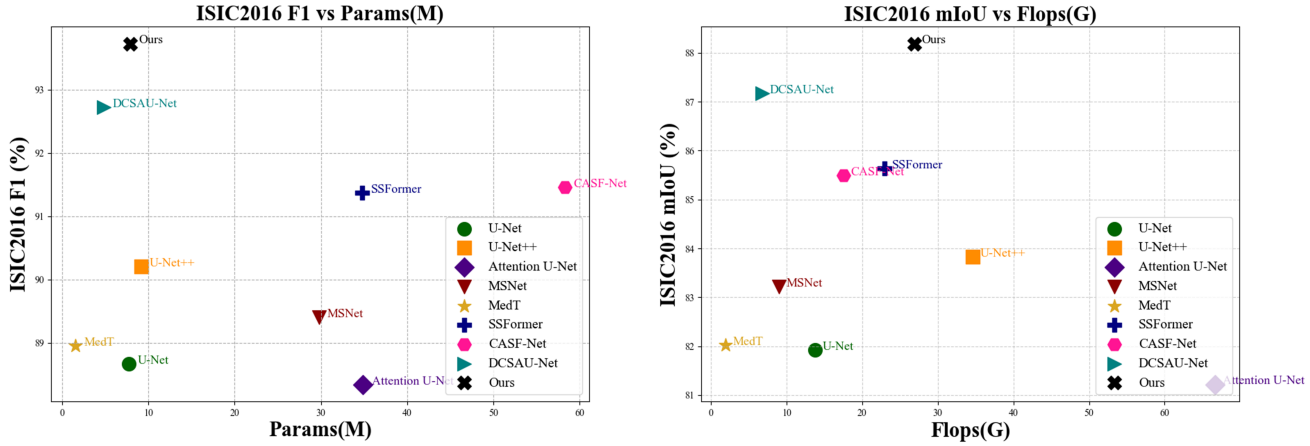

Figure 1. Performance comparison of AGTCNet with other models.

resented by AlexNet[3], have achieved significant success in image classification. However, classification tasks only need to recognize image categories without involving object boundaries or pixel-level distinctions. Fully Convolutional Networks (FCNs)[4] extend image prediction to the pixel level for the first time, achieving a key breakthrough from image classification to image segmentation.

However, excessive attention to pixel-level information may lead to the loss of boundary details. U-Net[5] partially mitigates this problem by introducing jump connections in the encoder-decoder structure, which preserves more boundary information. Based on its simplicity and scalability, several improved models have been derived. U-Net++[6] incorporates a Dense structure to fuse features of different resolutions and bridge the semantic differences in the convolutional layers; MALUNet[7] enhances feature interactions at different stages through a bridge attention module; and EGE-UNet[8] employs dilation convolution[9] to integrate multiscale features, realizing the combination of global and local information. However, CNNs are difficult to capture long-distance dependencies due to the sensory field limitation. Transformer[10], on the other hand, utilizes the self-attention mechanism[11, 12] to effectively solve this problem by dynamically adjusting the attention weights, which significantly enhances the ability to capture global information. Taking TransUnet[13] as an example, the model combines the global modeling capability of Transformer and the local information recovery of U-Net and demonstrates superior performance in multi-organ and heart segmentation tasks.

However, the pure Transformer model[14, 15] has limitations in capturing local details and lacks the translation invariance and local correlation of CNN, which leads to its underperformance in processing low-level information. To address this problem, hybrid models[16] combine the advantages of CNN and Transformer in vision tasks. TransFuse[17] significantly improves detail capture and overall characterization by parallel fusion of the local feature extraction capability of CNN with the global modeling capability of Transformer. Inspired by this, we propose a dual backbone network architecture, AGTCNet, which combines the detailed extraction of CNN and the global modeling of Transformer to achieve efficient feature representation and learning performance.

Furthermore, we note that some Transformer-based architectures use a fixed convolutional kernel for fine-grained feature extraction before operating on the attention mechanism, which limits the ability to capture multi-scale features. For this reason, we optimize this approach to better learn and characterize multi-scale features by dividing the input features into multiple channel groups and introducing depth-separable dynamic convolution, which allows the network to dynamically adjust the sensory fields according to the different scales of the image.

The contributions of this paper are as follows:

- We propose an Adaptive Grouped Transformer or AGT. This module efficiently combines adaptive grouped convolution and the Transformer architecture to achieve adaptive learning of multi-scale features.
- We propose the Curvature-based Guidance Enhancement Module or CGEM. This module directs the model to focus on the salient regions at the feature edges by capturing the curvature information of the feature maps.
- We construct a hybrid network based on AGT and curvature information for skin lesion detection, i.e., AGTCNet. Through extensive experimental validation on three publicly available skin lesion datasets, the re-

sults show that AGTCNet exhibits significant competitive advantages in several key performance metrics.

2. The proposed method

2.1. Adaptive Grouped Transformer

2.1.1 Patch Embedding Layer.

For a layered representation, we apply a patch embedding layer to adjust the properties of intermediate features. This layer specifically consists of two key steps: a 2 × 2 stepwise convolution as well as a normalization layer composition. The patch embedding layer is used to tune the feature scales as well as the channel dimensions, enabling the network to generate multi-scale feature representations at different stages.

2.1.2 Local Perception Module

In vision tasks, the absolute positional encoding used by the Transformer model introduces unique positional information for each image patch, thus potentially destroying translational invariance and increasing the instability of the model, making it difficult to generalize to new data with translational variations. To address this challenge, we introduce a local-awareness module. In this module, we borrow the idea of Res2Net[18], adopt a multi-scale processing approach, and develop the hierarchical representation into a single-block implementation. As shown in Fig.2, we group the feature images along the channel axis at different stages, and the number of channels in each group is $\frac{C}{S}$, where C represents the number of input feature image channels, S represents the number of groups in the group, and S takes the value of S $\in \{1, 2, 3, 4\}$. The specific implementation scheme is to partition the input tensor along the channel axis into S subsets, each of which has the shape of H×W×$\frac{C}{S}$, and then apply a 3 × 3 deep convolutional process to each feature subset, and finally, through the Concatenation operation to integrate the feature subsets into one feature representation. This strategy significantly enhances the spatial awareness of the output feature representation, making it more flexible and adaptive, while significantly reducing the computational complexity. The adaptive grouped convolution module can be defined as:

$$ACT\left(X\right) = Cat\left(DW_3\left(X_1\right), \cdots, DW_3\left(X_s\right)\right), s\epsilon\left\{1, 4\right\} \tag{1}$$

Where X denotes the feature input from the previous stage, $X \in \mathbb{R}^{H\times W\times C}$, $H \times W$ is the resolution of the input in the current stage, and C denotes the dimensionality of the features. Cat stands for Concatenation, DW_3 denotes deep convolution with a convolution kernel of 3, e denotes the Sth subset of features, and $X_S \in \mathbb{R}^{H\times W\times \frac{C}{S}}$, depending on the stage S takes the value $\{1, 2, 3, 4\}$.

2.1.3 Lightweight Multihead Self-Attention Module

To improve the stability of the input distribution during forward propagation and the stability of the gradient in backpropagation, we first perform LayerNorm[19] on the output of the feature from the LPM, with the normalization operation performed on the hidden dimension. Subsequently, the normalized features are fed into the Lightweight Multi-Headed Self-Attention Module (LMHSA). Different from the traditional self-attention mechanism, we reduce the computational complexity effectively by processing the deep convolution of K × K convolution kernel to reduce the spatial dimensions of K and V before computing the attention weights. Next, the combination of Q and K is utilized to compute the attention weights, which are applied to V to generate the weighted feature output.

In addition, to enhance the model's ability to model the relative positional relationships between elements in sequence data, we introduce a relative positional bias for each self-attention module B. The core process of the lightweight multi-head self-attention module can be defined as follows:

$$K' = DW_{k\times k}\left(K\right) \in \mathbb{R}^{\frac{n}{k^2}\times d_k} \tag{2}$$

$$V' = DW_{k\times k}\left(V\right) \in \mathbb{R}^{\frac{n}{k^2}\times d_k} \tag{3}$$

$$LightAttn\left(Q, K, V\right) = Softmax\left(\frac{QK'^T}{\sqrt{d_k}} + B\right)V'. \tag{4}$$

After that, according to the number of input heads h, h sequences of size h × d are generated and these sequences are connected into a comprehensive sequence of n × d to integrate the information of each attention head to form a more comprehensive feature representation. Next, the generated sequence information is normalized through the LayerNorm layer, and the processed feature information is directed to the IRFFN.

2.1.4 Inverse Residual Feedforward Network

As shown in Fig.2, the structure of IRFFN consists of two 1 × 1 convolutional layers for extending the feature dimensions and projecting to a lower dimensional feature space, respectively. To realize deeper feature transformations, we introduce a 3 × 3 deep convolutional layer between these two convolutional layers. The expanded features are processed by the GeLU[20] activation function and BatchNorm to further enhance the expressive power of the model. With

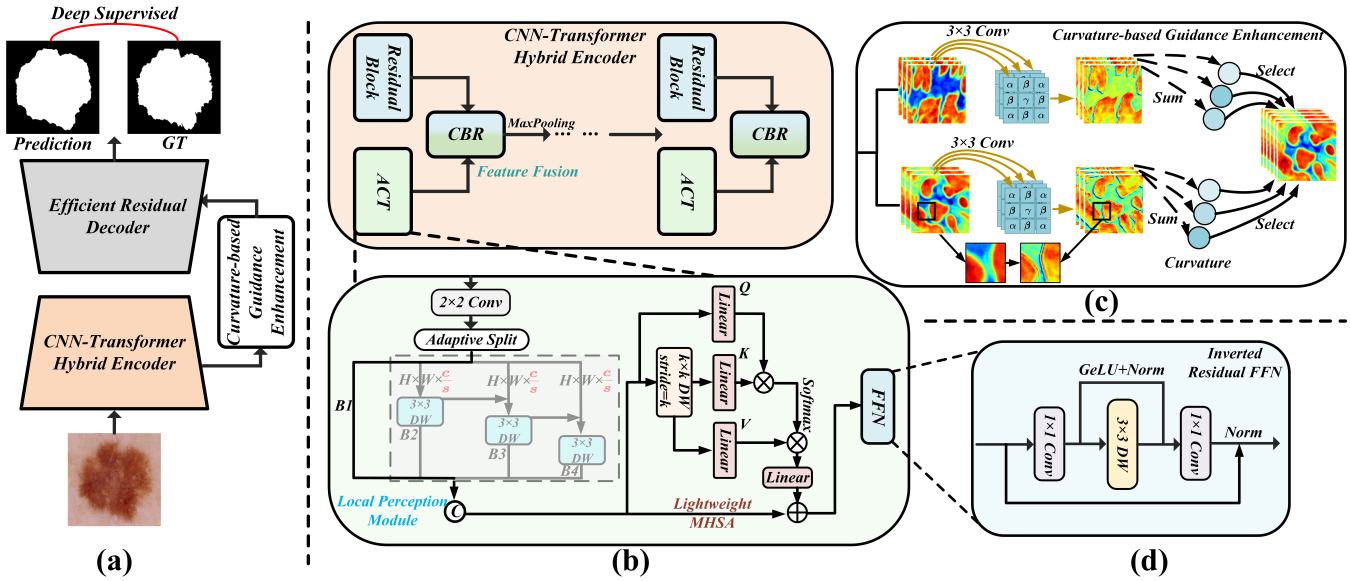

Figure 2. Hybrid network based on AGT and curvature information for skin lesion detection. (a) shows the main framework architecture of AGTCNet. (b) shows the detailed architecture of the AGT module, where (d) shows the inverse residual feedforward network (IRFFN) structure in the ACT module. (c) The figure shows the specific design of the CGEM module.

the introduction of residual structure, IRFFN effectively improves the propagation efficiency of gradient between different layers. Finally, the output features of IRFFN are normalized by BatchNorm to ensure the stability of the output features in a statistical distribution. The mathematical expression of IRFFN is as follows:

$$IRFFN\left(X\right)=Conv_1\left(DW_3\left(Conv_1\left(X\right)\right)\right) \quad (5)$$

Where X represents the feature sequence output by LMHSA, $Conv_1$ represents the convolution with convolution kernel 1, and DW_3 represents the depth-separated convolution with convolution kernel 3.

The above four modules constitute our proposed ACT module, which is mathematically represented as:

$$P_i=PE\left(X_{i-1}\right) \quad (6)$$

$$L_i=LPM\left(P_i\right) \quad (7)$$

$$Y_i=LMHSA\left(L_i\right)+L_i \quad (8)$$

$$X_i=IRFFN\left(Y_i\right)+Y_i \quad (9)$$

Where P_i, L_i, and Y_i denote the output characteristics of the PE, LPM, and LMHSA modules of the ith block, respectively.

2.2. Curvature-based Guidance Enhancement Module

To effectively deal with the problem of edge ambiguity in skin lesion regions, we propose a curvature-based guidance enhancement module (CGEM), whose detailed architecture is shown in Fig.2. In terms of curvature feature selection, we adopt the mean curvature as the main feature parameter because it can reflect the non-uniformity on the image surface more accurately, which helps to improve the model's ability in edge feature capture. By borrowing the simplified linear convolutional computation method proposed by Gong et al.[21]. we can efficiently approximate the mean curvature solution, which enhances the model's performance in recognizing and segmenting edge region features. The relevant formulas are as follows:

$$C=\left[C_1\ C_2\ C_3\right]\circledast X \quad (10)$$

Where the values of $C_1=\left[\alpha,\beta,\alpha\right]^T$, $C_2=\left[\beta,\gamma,\beta\right]^T$, $C_3=\left[\alpha,\beta,\alpha\right]^T$, α, β, and γ are -1/16, 5/16 and -1, respectively. $\circledast$ denotes convolution, X denotes the input image and C denotes the mean curvature.

2.3. AGTCNet

Medical images usually contain multimodal information, such as morphological and textual information. Convolutional neural networks excel at extracting both low-level and high-level visual features from images, while the Transformer model performs well in processing sequence data as well as linguistic information while being able to capture high-level semantic information in images. In this paper,

we construct a hybrid network based on AGT and curvature information for skin lesion detection. The model significantly improves the segmentation accuracy while maintaining high efficiency. The specific model parameters (Params) and floating point operations (FLOPs) are shown in Fig.1.

The specific structure of AGTCNet still utilizes a U-Net-like encoder-decoder architecture. In the encoder stage, we design a two-branch feature extraction backbone network fusing CNN and Transformer to obtain multimodal feature representations. The CNN branch employs simple residual blocks to focus on capturing local features. The Transformer branch generates multi-scale feature maps through a hierarchical cascade structure by stacking different numbers of ACT modules at each stage. In the first to fourth stages, 3, 3, 16, and 3 ACT modules are stacked, respectively, and the feature information extracted in each stage is retained and passed layer by layer to ensure multi-scale feature fusion.

To achieve the dual-branch interaction, we employ a simplified Convolution-Batch Normalization-Activation (CBR) with MaxPooling operation at each stage to facilitate the feature fusion between the CNN and Transformer branches. Meanwhile, a CGEM module is introduced between the encoder and decoder to strengthen the model's ability to learn and extract edge features. In addition, to supervise the feature reconstruction process, the network employs a deep supervision mechanism to compute the loss at different stages, which strengthens the training process by optimizing the loss. The mathematical representation of the loss function is as follows:

$$l_i = BCE\left(y, \hat{y}\right) + Dice\left(y, \hat{y}\right) \tag{11}$$

$$\mathbb{L} = \sum_{i=0}^{4} \lambda_i \times l_i \tag{12}$$

Where BCE and Dice represent binary cross-entropy loss and Dice loss, respectively. λ_i denotes the weights of the different stages. In the network, the values of λ_i 's at each stage are 0.4, 0.3, 0.2, 0.1.

3. Experiments

3.1. Experimental Parameters

All experiments were conducted in the Ubuntu 18.04 operating system and completed in the PyTorch 1.7.1 environment. Computational resources are provided by NVIDIA TITAN RTX to support efficient computation during training. Through extensive experimental validation, we set the initial learning rate of the model to 5e-4, the maximum number of training rounds to 400, and saved the optimal model and the latest rounds of the model during the training process. To enhance the generalization ability of the model, we introduce a variety of data enhancement techniques, including vertical flipping, horizontal flipping, and random rotation, to increase the diversity of samples and the robustness of the model.

Table 1. Performance metrics results for the various comparison methods on the ISIC2016 dataset. The best results are marked in red and the second best results are marked in blue (%).

Method	F1	mIoU	Precision	Recall
U-Net	88.66	81.92	90.48	91.17
U-Net++	90.20	83.83	92.94	90.68
Attention U-Net	88.34	81.22	**93.77**	87.32
MSNet	89.40	83.22	91.54	91.52
MedT	88.95	82.03	90.11	91.78
SSformer	91.37	85.63	90.18	93.22
CASF-Net	91.46	85.50	92.26	88.22
DCSAU-Net	**92.72**	**87.18**	91.42	**94.05**
Ours	**94.09**	**88.77**	**93.49**	**94.70**

3.2. Comparative Experiments

Fig.3 shows the segmentation results of representative samples in the ISIC2016 test set. It can be seen that the Ours method significantly outperforms the other methods in the segmentation of the overall lesion contour and local detail capture, and especially exhibits stronger performance when dealing with blurred regions. For sample A, although U-Net, MSNet[22], and CASF-Net can separate lesion regions with similar colors to the environment, Ours performs more finely in capturing edge details. For samples B and C, due to their complex edge features, although DCSAU-Net[23] and Ours can accurately capture the overall contours, Ours exhibits higher edge detection accuracy and robustness when dealing with complex edges (e.g., the lower right of sample B and the upper right of sample C), which further proves its superiority in dealing with complex lesion edges.

Fig.4 demonstrates the segmentation results of the Ours method with other methods on the ISIC2017 dataset. In sample A, the previous method misidentifies the lesion region as background, resulting in significant deviation from GT, which may stem from insufficient capture of complex features; Ours accurately identifies the lesion region by enhancing the capture of contextual information, with delicate edge processing, and the result is more closely aligned with GT. Sample C is subjected to the interference of hair, which generates obvious noise by methods such as U-Net.

In contrast, Ours effectively reduces the noise by combining local and global features through the two-branch backbone network. Sample D, the other methods failed to accurately reduce the noise, and the result is highly consistent with GT. In sample D, other methods fail to accurately segment the fuzzy lesion area, while Ours is closer to GT

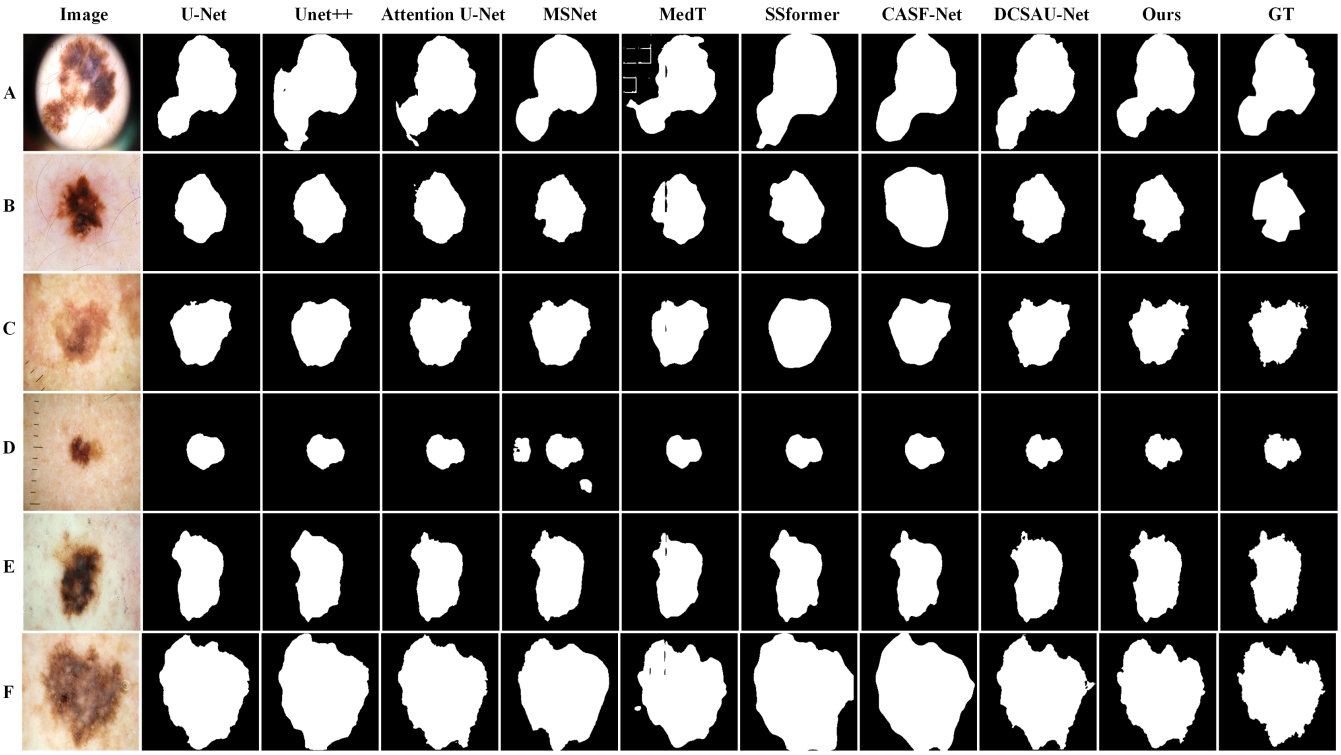

Figure 3. Visual presentation of predictions for selected samples from ISIC2016.

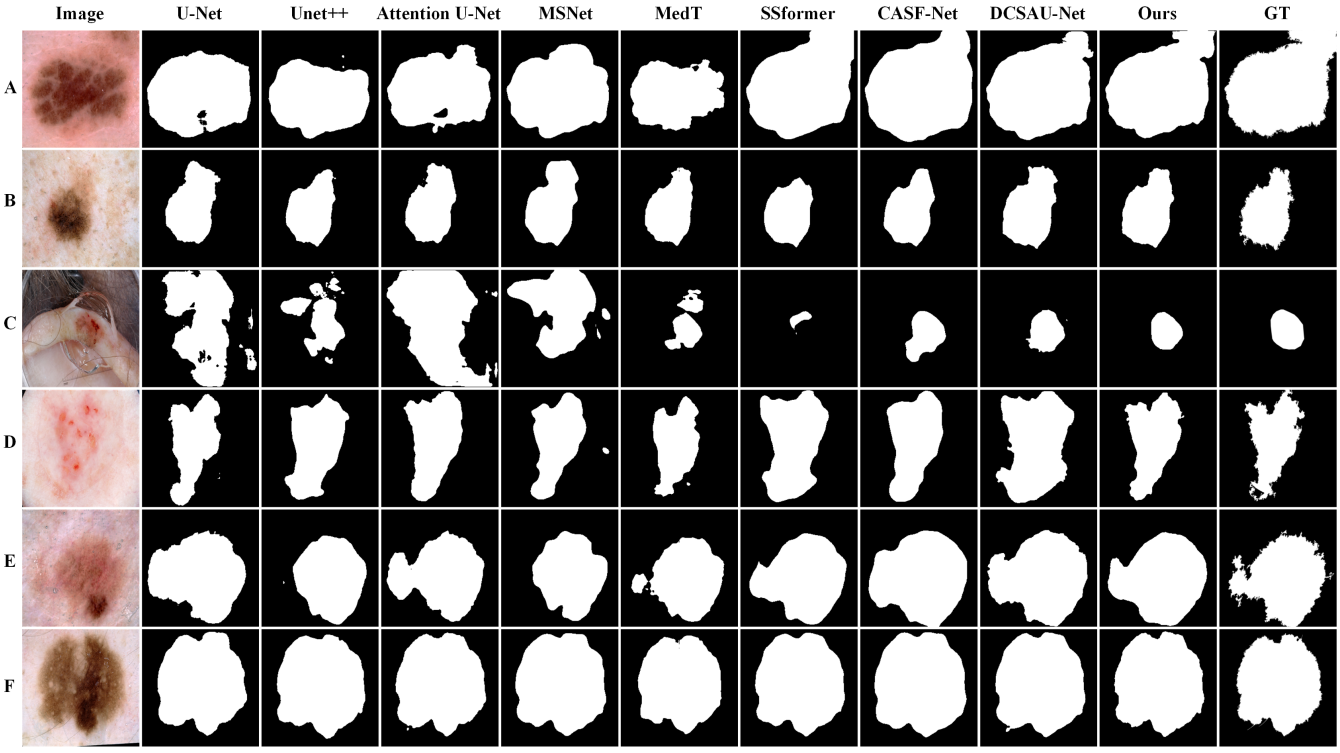

Figure 4. Visual presentation of predictions for selected samples from ISIC2017.

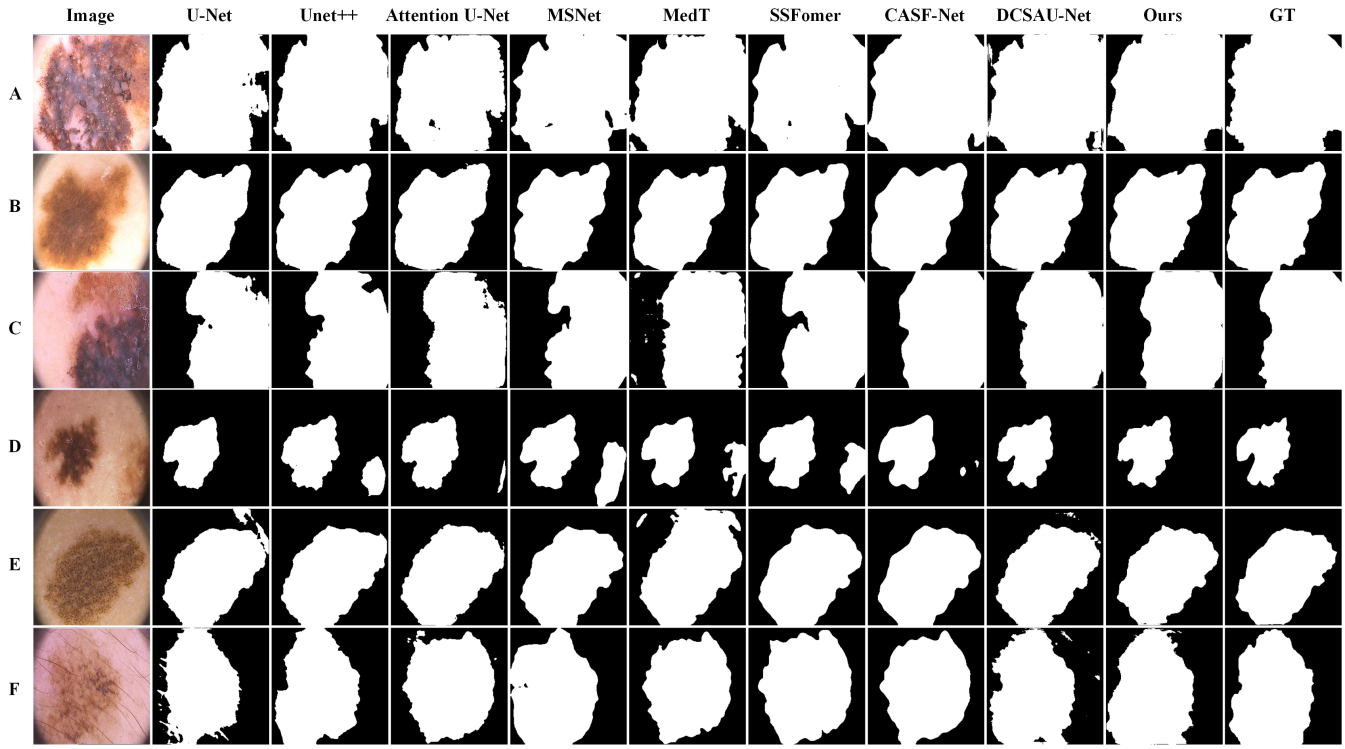

Figure 5. Visual presentation of predictions for selected samples in PH2.

Table 2. Performance metrics of the different compared methods on the ISIC2017 dataset. The best results are marked in red and the second best results are marked in blue (%).

Method	F1	mIoU	Precision	Recall
U-Net	78.73	64.92	89.66	70.18
U-Net++	80.16	66.89	**92.54**	70.70
Attention U-Net	80.66	67.59	88.15	74.34
MSNet	83.31	71.41	**91.07**	76.79
MedT	73.99	58.72	88.15	68.13
SSformer	83.43	71.30	81.51	85.54
CASF-Net	84.20	72.71	85.14	84.51
DCSAU-Net	**85.93**	**75.71**	83.93	**88.01**
Ours	**87.36**	**77.49**	82.95	**92.26**

Table 3. Performance metrics of the different comparison methods on the PH2 dataset. The best results are marked in red and the second best results are marked in blue (%).

Method	F1	mIoU	Precision	Recall
U-Net	86.43	76.21	84.43	88.51
U-Net++	93.57	88.07	93.40	93.74
Attention U-Net	92.35	85.93	91.66	93.05
MSNet	94.55	89.82	94.93	**94.18**
MedT	92.18	85.65	92.70	91.67
SSformer	93.12	87.28	94.53	91.76
CASF-Net	**94.60**	**90.08**	**95.44**	93.78
DCSAU-Net	94.11	89.03	95.00	93.23
Ours	**95.12**	**90.69**	**96.05**	**94.20**

by its stronger feature extraction ability.

Fig.5 illustrates the segmentation prediction visualization of the different methods on some samples of the PH2 dataset. The local zoom analysis of samples A, B, and C reveals subtle differences between SSFormer, CASF-Net, DCSAU-Net, and Ours. In Sample D, U-Net++ and Attention U-Net misidentify the background as a lesion, while U-Net and DCSAU-Net avoid this problem but are insufficient in capturing the edge details; Ours, on the other hand, significantly improves the overlap with GT through accurate edge processing.

In Sample E, the boundary processing of U-Net and MedT is confusing, while Ours achieves more accurate edge recognition by introducing curvature information. The segmentation results of Sample F show that the other methods have irregular boundaries in fuzzy regions, while Ours excels in accuracy and consistency and is closer to GT. Tab.1, Tab.2, and Tab.3 respectively summarize the performance of each method on the ISIC2016, ISIC2017, and PH2 datasets. AGTCNet achieved the best results in terms of F1 score, mIoU, precision, and recall, demonstrating its superior segmentation accuracy.

Table 4. Structural demonstration of different ablation variants.

Method	ACT	CGEM	Deep supervision	ISIC2016		ISIC2017		PH2	
				F1	**IoU**	**F1**	**IoU**	**F1**	**IoU**
AGTCNet-1	×	✓	✓	92.63	86.72	85.75	75.11	93.83	88.39
AGTCNet-2	✓	×	✓	93.39	87.65	86.55	76.17	94.58	89.72
AGTCNet-3	✓	✓	×	93.80	88.27	86.92	76.79	94.76	90.11
AGTCNet	✓	✓	✓	94.09	88.77	87.36	77.49	95.12	90.69

3.3. Ablation Experiments

In this section, we validate the effectiveness of the proposed module through ablation experiments, all of which are conducted based on the ISIC2016, ISIC2017 and PH2 datasets. Specifically, Tab.4 shows the different model variants. AGTCNet-1 removes the ACT module and uses only the Residual Block as the backbone structure to evaluate the role of the ACT module in global feature capture.

According to Tab.4, the three ablation variants have significant gaps with the AGTCNet model in terms of F1 and mIoU metrics, indicating that the ACT module, CGEM module, and the depth supervision mechanism play a key role in enhancing edge feature extraction, improving segmentation accuracy, and dealing with fuzzy regions robustly, verifying their effectiveness.

In the second set of samples, the edge processing results reveal the degree of misclassification of the models, especially AGTCNet-1 incorrectly recognizes the surrounding environment as the lesion region, validating the importance of the ACT module in contextual understanding and separation of lesion and background. In addition, the performance metrics in Tab.4 further support the effectiveness and robustness of AGTCNet in fuzzy region processing and edge detail recognition.

This conclusion is further supported by the data in Tab.4, which shows that the overall performance of the three ablation models is lower than that of AGTCNet, validating the effectiveness of the proposed module.

4. Conclusion

In this paper, a novel hybrid network named AGTCNet is proposed for skin lesion detection and targeted to solve the problem of incomplete extraction of the edge region of skin lesions. AGTCNet adopts a two-branch backbone network structure based on CNN and Transformer, which is able to capture both global features and local detail information, thus significantly enhancing the network's characterization ability. In order to improve the model's deficiency in edge detail processing, a CGEM module is designed to guide the model to accurately recognize lesion edges by collecting curvature information. In addition, AGTCNet introduces a deep supervision mechanism to dynamically supervise the feature loss and optimize the learning strategy in real time based on feedback. The effectiveness and superiority of the AGTCNet network architecture is verified through extensive experiments on three public datasets.

References

[1] Zhiwei Dong, Genji Yuan, Zhen Hua, and Jinjiang Li. Diffusion model-based text-guided enhancement network for medical image segmentation. *Expert Systems with Applications*, 249:123549, 2024. 1

[2] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241, 2015. 2

[6] 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: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11, 2018. 2

[7] 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, 2022. 2

[8] 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, 2023. 2

[9] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[11] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational Visual Media*, 9(4):733–752, 2023. 2

[12] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368, 2022. 2

[13] 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. 2

[14] Yue He, Lan Chen, Yu-Jie Yuan, Shu-Yu Chen, and Lin Gao. Multi-level patch transformer for style transfer with single reference image. In *International Conference on Computational Visual Media*, pages 221–239, 2024. 2

[15] Zhongqi Wu, Jianwei Guo, Chuanqing Zhuang, Jun Xiao, Dong-Ming Yan, and Xiaopeng Zhang. Joint specular highlight detection and removal in single images via unet-transformer. *Computational Visual Media*, 9(1):141–154, 2023. 2

[16] Zhiwei Dong, Jinjiang Li, and Zhen Hua. Transformer-based multi-attention hybrid networks for skin lesion segmentation. *Expert Systems with Applications*, 244:123016, 2024. 2

[17] Y Zhang, H Liu, and Q TransFuse Hu. Fusing transformers and cnns for medical image segmentation. arxiv 2021. *arXiv preprint arXiv:2102.08005*. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645, 2016. 3

[19] Jimmy Lei Ba. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[20] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[21] Yuanhao Gong and Ivo F Sbalzarini. Curvature filters efficiently reduce certain variational energies. *IEEE Transactions on Image Processing*, 26(4):1786–1798, 2017. 4

[22] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic polyp segmentation via multi-scale subtraction network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 120–130, 2021. 5

[23] Qing Xu, Zhicheng Ma, HE Na, and Wenting Duan. Dcsau-net: A deeper and more compact split-attention u-net for medical image segmentation. *Computers in Biology and Medicine*, 154:106626, 2023. 5