This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3463713

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

TrUNet: Dual-branch network by fusing CNN and Transformer for skin lesion segmentation

Wei Chen[1,3], Qian Mu[1], and Jie Qi[2]
[1]School of Communication and Information Engineering, Xi'an University of Science and Technology, Xi'an, Shaanxi, 710054, China
[2]Orthopedic Department, Shaanxi Provincial People's Hospital, Xi'an, Shaanxi, 710068, China
[3]Xi'an Key Laboratory of Network Convergence Communication, Xi'an, Shaanxi, 710054, China

Corresponding author: Qian Mu (e-mail: 22207223135@stu.xust.edu.cn).

ABSTRACT In the medical field, precise segmentation of skin lesion areas is essential for accurate diagnosis and treatment of diseases. Due to the varied morphologies and fuzzy boundaries of skin lesions, as well as interference from hair coverage, segmentation tasks are extremely challenging. To address the problem, a network called TrUNet is proposed, which combines the advantages of Transformer and convolutional neural networks (CNNs). Transformer and Res2Net are taken as two branches of the encoder in this network, with the goal of extracting rich global information for precise lesion segmentation in medical images. Firstly, the TrFusion module was designed to selectively fuse complementary features extracted by the Transformer branch and the Res2Net branch in the encoder, enhancing important information while suppressing irrelevant details. Secondly, the Multi-Scale Feature Aggregation (MFA) module was designed to fuse feature representations from different stages of the same branch to complement positional and spatial information. Finally, to validate the effectiveness of the proposed method, experiments were conducted on the ISIC2017, ISIC2018, and PH2 datasets. TrUNet achieved Dice coefficient of 90.61%, IoU of 84.25%, and Accuracy of 94.74% on the ISIC2018 dataset. This indicates that our model has enormous potential in the field of medical image segmentation.

INDEX TERMS Deep learning, convolutional neural network, transformer, skin lesion segmentation

I. INTRODUCTION

The skin is one of the largest organs in the human body and a critical line of defense for the immune system. Exposed to the external environment, the skin is vulnerable to factors such as ultraviolet (UV) radiation, temperature changes, and pathogens, which increase the risk of diseases. Melanoma is a fatal skin cancer [1], with early-stage cure rates as high as 90%. However, once melanoma cells spread through the body's circulatory system into other normal tissues in late stages, the cure rate drops to only 10%. Therefore, early diagnosis and treatment are essential.

In clinical diagnosis, dermatoscopy is commonly used to generate high-resolution images of skin lesions for lesion area segmentation. However, lesion areas often have blurry boundaries, low contrast, and can be obscured by hair. Traditional manual segmentation methods are slow, labor-intensive, and prone to subjective biases. Therefore, it is particularly important to develop automated skin lesion segmentation techniques.

Computer vision and deep learning algorithms can automatically recognize and accurately delineate abnormal regions in skin images. This technology is crucial in clinical practice because precise segmentation of skin lesions enables doctors to identify and locate affected areas with greater accuracy. This capability supports early diagnosis and treatment, leading to improved recovery rates and enhanced quality of life for patients.

Deep learning is a machine learning method that can imitate the information transmission process between neurons in the human brain to learn the representation and characteristics of data. With the rapid development of deep learning, CNNs is also widely used in the field of medical image segmentation. Researchers have also proposed various algorithms to solve the problems of blurred edge of skin lesions and low segmentation accuracy. Long et al. [2] proposed the Fully Convolutional Network (FCN), which was the first network to introduce a fully convolutional structure for image segmentation, pioneering the field of semantic segmentation. Ronneberger et al. [3] proposed U-Net, a U-shaped network with an encoder-decoder architecture. This network connects the encoder to the decoder through skip connections, enabling the fusion of high-level and low-level semantic features. Subsequently,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3463713

networks based on the U-shaped structure were proposed, such as UNet++ [4], U2-Net [5], UNet3+ [6], Res-UNet [7], and Dense U-Net [8], among others.

In addition, in the field of object detection, researchers have proposed a feature disentanglement module to address the inherent feature misalignment between classification and regression tasks [9]. This method disentangles features in the Feature Pyramid Network (FPN) and reduces inconsistent responses and suppresses inferior predictions through a response alignment strategy. These methods significantly improve the performance of object detection and demonstrate the effectiveness of feature disentanglement in complex tasks.

However, the context of medical images is complex, so it is necessary to effectively extract and utilize contextual feature information at multiple scales. Recently, researchers have proposed methods to integrate multi-scale information, such as PsPNet [10], DeepLabV3+ [11], CE-Net [12] and HrNet [13], among others. These networks capture rich contextual information at different scales, which helps improve the model's ability to recognize and segment objects at different scales. Despite their powerful feature extraction capabilities, methods based on CNNs are unable to capture long-distance dependency information due to the limitations of the convolution operation itself. As a result, it is less effective in processing images with significant structural differences.

Vaswani et al. [14] proposed the Transformer model, which is good at modeling global context and has limitations in capturing fine-grained details, whereas CNN is good at capturing local features in an image. Therefore, both feature extraction functions complement each other. Some recent studies have combined CNN and Transformer for medical segmentation. TransUNet proposed by Chen et al. [15] and Swin-Unet proposed by Cao et al. [16], along with subsequent research, have achieved breakthroughs in segmentation effectiveness compared to previous algorithms. However, shallow networks frequently underutilize the copious spatial information inherent in their data. They typically confine their contextual modeling to a singular scale, thereby disregarding the interrelated dependencies and coherence that span various scales within the data. Wu et al. [17] proposed a HorUNet model with higher-order spatial interactions based on recursive gated convolution and added a multi-stage dimensional fusion mechanism to the skip connection part, resulting in the MHorUNet model architecture. This model exhibits high segmentation accuracy. Zhang et al. [18] proposed TransFuse, a network that combines Transformer and CNN in a parallel manner to efficiently capture both global dependencies and low-level spatial details in a shallower manner for medical image segmentation. The above method combines the models of Transformer and CNN to show significant advantages in medical image segmentation. Transformers excel in extracting global features and modeling long-range dependencies, while CNNs are skilled at accurately extracting local features and capturing fine-grained details. Their integration complements each other, resulting in a significant enhancement in segmentation accuracy and effectiveness.

The skin lesion areas exhibit characteristics such as uneven pixel distribution, significant morphological variations, and blurred edge contours. These features severely weaken the correlation among the lesion areas. During image segmentation, these issues lead to the loss of detailed information and mis-segmentation of the lesion areas. Inspired by the above studies, this paper proposes a novel network architecture for medical image segmentation called TrUNet, using the U-shaped architecture as a reference. TrUNet consists of two branches of Res2Net [19] and Transformer as encoders. The network utilizes the two branches of the encoder to extract rich global information.

In addition, it introduces MFA module and TrFusion module for complementary information integration and feature fusion.

The main contributions are summarized below:

(1) This paper introduces TrUNet, a novel dual-encoding medical segmentation framework. The network utilizes a dual-encoding architecture incorporating Res2Net and Transformer to extract both global and local features, establishing multiscale long-range dependencies without the need for deep hierarchical networks, effectively capturing global information.

(2) The MFA module is designed to fuse features from different stages of the same branch, supplementing positional and spatial information. The TrFusion module selectively integrates features from different branches at the same stage to enhance important information while suppressing irrelevant details.

(3) To validate the effectiveness and generalization capability of the network, experiments were conducted on skin lesion datasets ISIC2017, ISIC2018, and PH2, comparing them with currently popular methods. The experimental results clearly show that the proposed algorithm outperforms other state-of-the-art (SOTA) algorithms, indicating its superior performance compared to existing methods.

The main research of this paper is as follows: Section 2 will introduce the SOTA of medical image segmentation research. Section 3 describes in detail the methodology used in this paper. Section 4 presents experiments. Section 5 provides conclusions.

II. RELATED WORK

A. CNN IN MEDICAL IMAGE SEGMENTATION

In recent years, deep learning methods based on CNNs have been widely applied in the field of medical image segmentation [20-22]. Researchers have also proposed various algorithms to address issues such as fuzzy edges of

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3463713

skin lesion regions and low segmentation accuracy. Yuan et al. [23] proposed a deep fully convolutional automatic skin lesion segmentation algorithm based on the Jaccard distance. He et al. [24] used chained residual pooling to capture contextual information and further improved the performance of skin lesion segmentation by integrating network with Conditional Random Field post-processing. Bi et al. [25] utilized Generative Adversarial Networks (GANs) for stacked adversarial learning of skin lesion features, enhancing the segmentation performance of FCN. Huang et al. [26] introduced an end-to-end object scale-oriented FCN (OSO-FCNs) for lesion segmentation. Berseth et al. [27] applied U-shaped networks in skin lesion segmentation. U-shaped networks based on encoder-decoder structures have become mainstream for segmentation tasks. As research progresses, improved versions of U-shaped networks continue to emerge. Tang et al. [28] developed a multi-stage UNet (MS-UNet) that integrates a deeply supervised learning strategy. They incorporated multiple U-Nets into the auto-context scheme to improve skin lesion segmentation. To better represent feature maps, Schlemper et al. [29] proposed the Attention U-Net, which introduces attention mechanisms that adaptively adjust feature weights for different spatial positions, thereby enhancing focus on important regions. Alom et al. [30] proposed R2U-Net, a model that adds residual and recurrent networks to U-Net to avoid the network being too deep to learn the gradient.

Improvements to U-Net include modifications to its encoder, decoder, and jump connections, but fail to address significant long-range dependencies between pixels.

B. TRANSFORMER IN MEDICAL IMAGE SEGMENTATION

The Transformer is a neural network architecture based on the self-attention mechanism, widely used in the field of Natural Language Processing (NLP). The core idea is to utilize self-attention mechanism to process sequential data, taking into account information from different positions in the sequence, thus avoiding the limitations of local receptive fields in CNNs.

Researchers combined knowledge from computer vision (CV) and NLP fields to apply the Transformer architecture with global attention mechanism to full-sized images, leading to the development of the Vision Transformer (ViT) [31]. Song et al. [32] proposed a TGDAUNet network consisting of a dual-branch backbone of CNNs and Transformers and a parallel attention mechanism, and explored the potential semantic relationships between boundaries and regions to further refine the target boundaries. Chen et al. [33] proposed the CoTrFuse network, which consists of EfficientNet and Swin Transformer [34] architectures, to improve the performance of medical image segmentation through the use of skip connections and feature fusion. Chen et al. [35] propose TransAttUNet, which combines Transformer's self-attention mechanism with convolutional global spatial attention for semantic segmentation tasks. Lin et al. [36] proposed DS-TransUNet, a Swin Transformer based model for medical image segmentation tasks that incorporates parallel dual-scale encoding and a Transformer Interactive Fusion module for complementary encoding information across different scale patches.

The introduction of these methods has enriched research in the field of medical image segmentation, providing novel insights and technical means for medical image processing. It holds promise for achieving better outcomes in clinical practice.

III. PROPOSED METHOD

In this section, the architecture of the TrUNet network will be described, providing detailed information about its constituent modules. The overall structure of the network, which includes key components such as the MFA module and TrFusion module, will be outlined first.

A. OVERALL STRUCTURE

CNNs are effective in capturing local features in the field of image processing, but they have limitations in capturing long-range dependencies. Conversely, Transformer models excel at modeling global contexts but perform poorly in capturing fine-grained details. Given the complementary nature of feature extraction between the two, recent research has started to combine CNNs and Transformers to enhance the effectiveness of medical image segmentation tasks.

Based on the above inspiration, this paper introduces TrUNet, a novel architecture for medical image segmentation. Firstly, TrUNet utilizes Res2Net and Transformer as two branches of the encoder to jointly extract multi-scale feature information and establish long-range dependencies. Secondly, the features extracted from the two branches are passed to the MFA module to generate fused high-level semantic feature information, thereby enhancing the network's understanding of complex scenes. Thirdly, features with the same resolution are input into the TrFusion module to effectively fuse feature information. Finally, by combining the fused feature map information, segmentation predictions are generated using the attention-gated (AG) skip connections. The architecture of the TrUNet network is shown in FIGURE 1.

The network proposed in this paper has the following advantages: (1) By employing two independent encoders to extract features from different perspectives, it enables the acquisition of rich and diversified feature information. (2) The interaction and fusion of information between the two encoders can establish an effective link between features at different levels and resolutions, enhancing the network's ability to utilize features across various scales and levels, while also enhancing the suppression of irrelevant details and highlighting important information. (3) It integrates feature representations from different stages of the same branch to generate fused high-level semantic feature information, without increasing model complexity or computational burden.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3463713

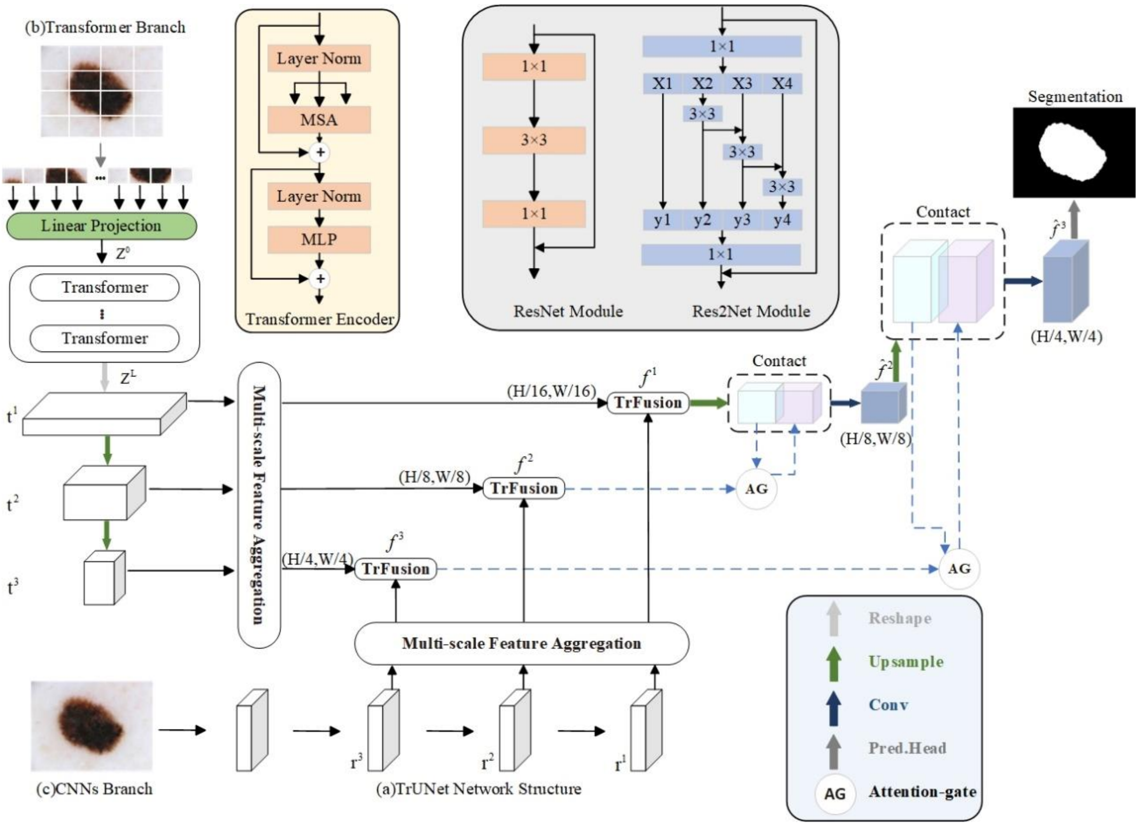

FIGURE 1. Illustration of the proposed TrUNet.

B. TRANSFORMER BRANCH

Transformer is a neural network architecture based on self-attention mechanism, primarily applied in the field of NLP. It processes sequence data through self-attention mechanism, possessing high parallelism and suitability for handling long sequences and capturing long-range dependencies. In image processing, Transformer effectively captures global information, enabling better extraction of complex structures and features from images.

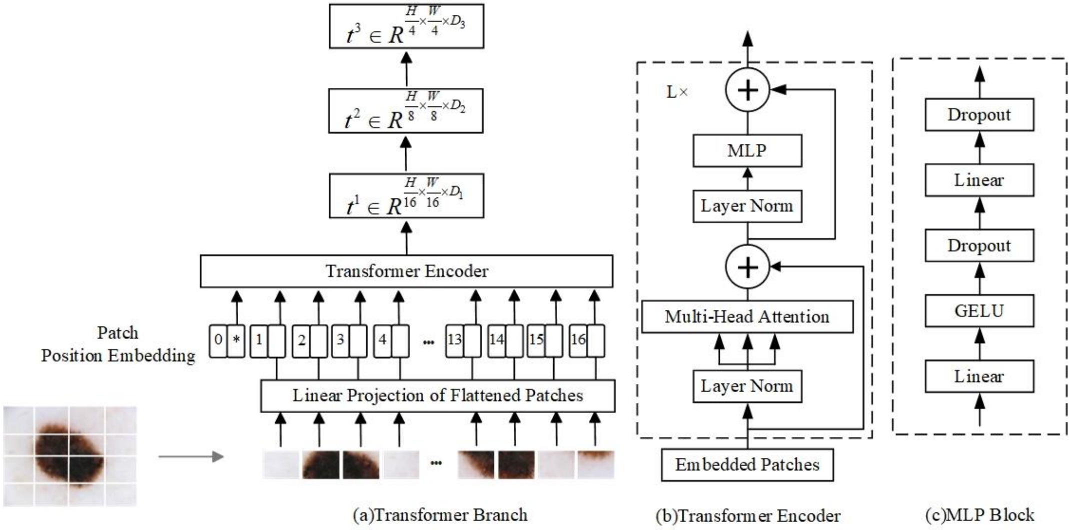

FIGURE 2. Illustration of the Transformer branch.

The design of the Transformer branch follows the typical encoder-decoder architecture. The Transformer Encoder consists of multiple layers of Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP). MSA is an extended form of Self-Attention, and the calculation of SA is as follows:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3463713

$$Attention(Q,K,V)=\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V \qquad (1)$$

In the formula, Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimension of the key vectors.

In this paper, firstly, the input image $x \in \mathbb{R}^{H\times W\times 3}$ is divided into $N=\frac{H}{S}\times\frac{W}{S}$ patches, where H represents the height of the input image, W represents the width, and S represents the side length of each image patch. Secondly, these patches are flattened and passed through a linear embedding layer with an output dimension of D_1 to generate the raw embedding sequence. Finally, add trainable positional embeddings of the same dimension as the original embeddings to the original embedding sequence. This results in the embedding vector $Z^0 \in \mathbb{R}^{N\times D_1}$. Z^0 is fed into the Transformer encoder to obtain the encoded sequence $Z^L \in \mathbb{R}^{N\times D_1}$. For the decoder part, Z^L is reshaped back to $t^1 \in R^{\frac{H}{16}\times\frac{W}{16}\times D_1}$ by progressive upsampling (PUP) method. Followed by two consecutive standard upsampling-convolution layers to recover to recover the spatial resolution, yielding $t^2 \in R^{\frac{H}{8}\times\frac{W}{8}\times D_2}$ and $t^3 \in R^{\frac{H}{4}\times\frac{W}{4}\times D_3}$. The feature maps at different scales t^1, t^2 and t^3 are saved and fused with the corresponding stage feature maps from the Res2Net branch. The architecture of the Transformer branch is shown in FIGURE 2.

C. CNN BRANCH

The core idea of Res2Net lies in enhancing the perceptual capabilities of neural networks by constructing multi-scale feature maps. The approach is realized by introducing a hierarchical level of parallel connections, containing multiple parallel sub-network modules inside each level. These sub-network modules process feature maps at different scales separately and integrate them together to represent multi-scale features at a finer granularity and expand the perceptual field of each network layer.

Considering the characteristics of Transformer, this study removes the Layer4 and FC layers from Res2Net and uses the Transformer branch to obtain global context information. A relatively shallow model is designed while retaining rich local information. Typically, Res2Net consists of five blocks, each of which downsamples the feature map twice. In this paper, the outputs of Layer1, Layer2 and Layer3 are extracted and fused with the outputs of t^1, t^2 and t^3 of the corresponding stages of Transformer respectively. This fusion strategy can effectively combine local details and global information to improve the performance of the model in complex tasks.

D. MFA MODULE

The morphological differences of skin lesions are notable, frequently leading to issues such as under-segmentation and over-segmentation, thereby complicating the segmentation process. To address this challenge, this paper proposes a MFA module, specifically tailored to effectively merge feature representations from various stages of the same branch, resulting in enhanced fused high-level semantic feature information.

In the MFA module, feature information from different stages of the same branch is fused with each other through upsampling and downsampling operations, and the information at different levels are effectively integrated to fully utilize the location information and spatial information to generate more comprehensive and rich high-level semantic feature information, and to improve the network's comprehension of the complex scene. The structure of the MFA module is shown in FIGURE 3.

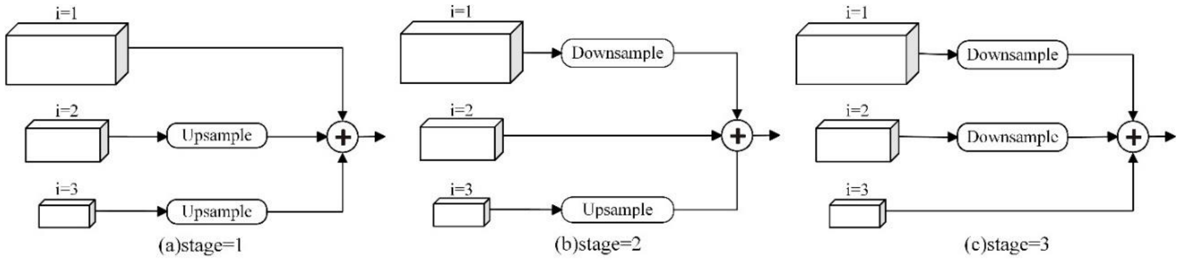

FIGURE 3. Illustration of the proposed MFA module.

E. TRFUSION MODULE

CNN excels at capturing local features and details in images, while Transformers are adept at handling global relationships and semantic information. To effectively combine the encoding advantages from both branches, the TrFusion module is proposed to mutually integrate features extracted from the two branches, enhancing important information and suppressing irrelevant details.In this paper, the Channel Attention mechanism SENet is employed to enhance the global information of the Transformer branch, while the Spatial Attention mechanism serves as a spatial filter to process features extracted by the CNN branch, enhancing local details and suppressing irrelevant regions. To better integrate complementary information from the two branches, a combination of Depthwise Separable

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3463713

Convolution (DSC) and Convolution is utilized, making communication between the two branches more efficient. The TrFusion module utilizes both channel attention and spatial attention to enhance global information while boosting local details, ultimately generating a fused feature representation to improve the network's understanding and representation capability of image data. The features extracted by the Transformer branch are denoted as t^i, those extracted by the CNN branch as r^i, and the fused features as f^i. The following equations represent the process of fusing features from the dual branches:

$$\hat{t}^i = ChannelAttention(t^i) \tag{2}$$

$$\hat{r}^i = SpatialAttention(r^i) \tag{3}$$

$$\hat{t}_1^i = DSConv(\hat{t}^i) \tag{4}$$

$$\hat{t}_2^i = Conv(\hat{t}^i) \tag{5}$$

$$\hat{r}_1^i = DSConv(\hat{r}^i) \tag{6}$$

$$\hat{r}_2^i = Conv(\hat{r}^i) \tag{7}$$

$$\hat{t}_f^i = \hat{t}_1^i \times \hat{r}_2^i \tag{8}$$

$$\hat{r}_f^i = \hat{t}_2^i \times \hat{r}_1^i \tag{9}$$

$$\hat{b}^i = Conv(t^i W_1^i \odot r^i W_2^i) \tag{10}$$

$$f^i = Residual([\hat{b}^i, \hat{t}_f^i, \hat{r}_f^i]) \tag{11}$$

$|\odot|$ is the Hadamard product and Conv is a 3×3 convolution layer. The Hadamard product models fine-grained interactions between features from two branches to obtain $\hat{b}^i$. $\hat{b}^i$ is concatenated with attention features $\hat{t}_f^i$ and $\hat{r}_f^i$ to form the fused feature f^i. The fused feature f^i effectively captures both global and local information at the current spatial resolution. The structure of the TrFusion module is shown in FIGURE 4.

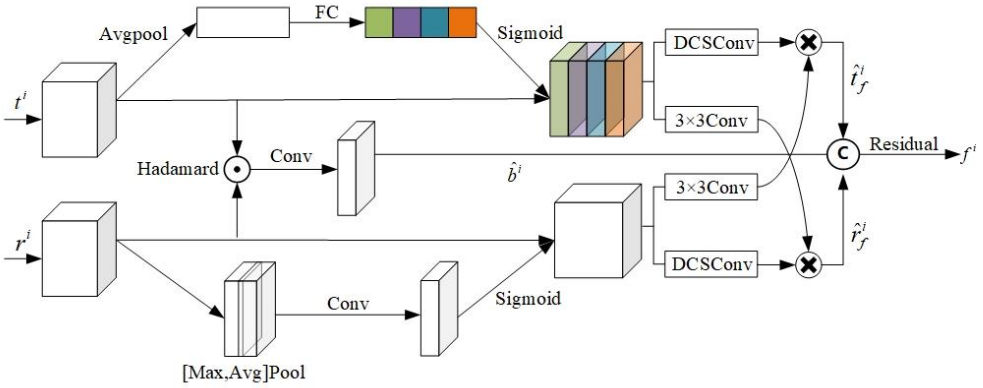

FIGURE 4. **Illustration of the proposed TrFusion module.**

IV. EXPERIMENT

In this section, specific details of the experiments are presented, including the dataset used, implementation details, and comparisons with currently popular methods as benchmarks. Additionally, ablation experiments are conducted to analyze the validity of the proposed model.

A. DATASET

To better evaluate the effectiveness of the network, this study utilizes three publicly available datasets: ISIC2017 [37], ISIC2018 [38,39], and PH2 [40], released by the International Skin Imaging Collaboration (ISIC) and by Mendonca et al., respectively, as experimental samples.

TABLE I
SUMMARY OF DATASET QUANTITIES

Datasets	Usage	Structure
ISIC2017	Training/Validation/Testing Ablation study	2000/150/600
ISIC2018	Training/Validation/Testing	2594/100/1000
PH2	External Testing	0/0/200

The ISIC2017 dataset comprises 2000 training images, 150 validation images, and 600 test images. The ISIC2018 dataset consists of 2594 training images, 100 validation images, and 1000 test images. The PH2 dataset includes 200 images of skin lesions, serving as an additional test set for the ISIC2017 dataset. PH2 is not involved in the model training process. The summary of the quantities of the three public datasets is shown in TABLE I.

B. DATA PREPROCESSING

The dataset is one of the important factors affecting the deep learning network, and in this paper, three datasets, ISIC2017, ISIC2018, and PH2, are used respectively. Skin lesion regions are all characterized by low contrast, fuzzy boundaries, and hair occlusion, which affect the segmentation accuracy. In order to improve the segmentation accuracy of the lesion region, this paper carries out preprocessing operations on images with the following steps:

(1) Hair removal: This paper utilizes morphological operations and black hat transformation to detect hair contours. Hair contours are enhanced through threshold operations, and the original image is restored based on these contours, resulting in the image after hair removal.

(2) Contrast enhancement: Custom adjustments of brightness and contrast are applied to adapt the image to application requirements. The mean and standard deviation of each image are computed and normalized. Subsequently, the normalized pixel values are mapped to a preset target range of

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

mean and standard deviation to enhance the contrast and brightness of the image.

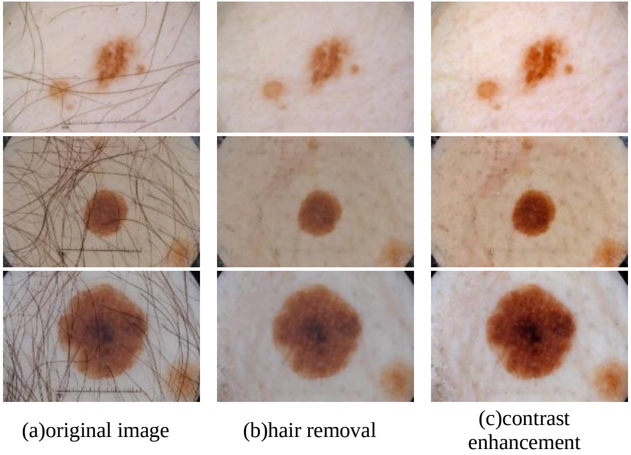

FIGURE 5. **Comparison of preprocessed skin lesion images.**

FIGURE 5 illustrates the contrast results of skin lesion images before and after preprocessing. (a) shows the original image, which is affected by hair occlusion and blurred boundaries of lesion areas. The image after hair removal in (b) largely addresses the issue of hair occlusion, yet the contrast remains low, and the boundaries remain blurred. (c) presents the result after contrast enhancement applied to the image in (b), where the contrast between lesion areas and surrounding regions is significantly enhanced, and there is no hair occlusion.

C. TRAINING SETTINGS AND COMPARISON METRICS

The experiments in this paper were conducted on the Ubuntu 20.04 operating system. The deep learning environment was set up with CUDA 11.3 and PyTorch 1.8.1 framework, utilizing the NVIDIA RTX 3090 GPU for accelerated model training. To enhance the model's generalization capability, data augmentation was performed on the ISIC2017 and ISIC20180 training datasets, including random translations, scaling, rotations, and flips. The input image size for all three datasets was standardized to 192×256 pixels. Processed images and labels were saved as NumPy array files for training purposes. The Adam optimizer was utilized with a learning rate of 1e-4, employing a patch size of 16×16. Model training iterated for 100 epochs with a batch size of 32.

Evaluation metrics are direct representations of measuring the segmentation effectiveness of a model. In this paper, Dice Similarity Coefficient (Dice), Intersection over Union (IoU), Accuracy, Recall, Precision, and Hausdorff distance (HD) are utilized as evaluation metrics for segmentation results, with calculation methods as follows:

$$Dice = \frac{2\times TP}{2\times TP+FN+FP} \tag{12}$$

$$IoU = \frac{TP}{TP+FN+FP} \tag{13}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{14}$$

$$Recall = \frac{TP}{TP+FN} \tag{15}$$

$$Precision = \frac{TP}{TP+FP} \tag{16}$$

$$HD(A,B) = \frac{1}{N+1}\sum_{i=0}^{N} \max(h(A,B),h(B,A)) \tag{17}$$

Where TP represents true positives, indicating the number of pixels correctly segmented within the lesion area. TN represents true negatives, indicating the number of pixels correctly segmented outside the lesion area. FP represents false positives, indicating the number of pixels incorrectly segmented within the lesion area. FN represents false negatives, indicating the number of pixels incorrectly segmented outside the lesion area.

D. EXPERIMENTAL RESULTS

Experiments were conducted on three datasets: ISIC2017, ISIC2018, and PH2, comparing the TrUNet model with six mainstream segmentation methods: U-Net, Deeplabv3+, PsPNet, HrNet, TransFuse, and MHorUNet. The segmentation performance of these methods was evaluated using metrics including Dice, IoU, Accuracy, Recall, Precision, and HD to validate the effectiveness of the TrUNet segmentation model.

1) RESULTS ON THE ISIC2017 DATASET

The experimental results of comparing the TrUNet model proposed in this paper with the current mainstream methods on the ISIC2017 dataset are shown in TABLE II. The Dice of TrUNet is 87.83%, the IoU is 80.27%, the Accuracy is 94.58%, the Recall is 88.88%, the Precision is 90.88% and the HD is 4.21. While TrUNet did not achieve the highest Precision and HD, it demonstrated the best performance across the remaining four evaluation metrics.

To verify the generalization of the model, this paper uses the PH2 dataset as an additional test set to the ISIC2017 dataset, and PH2 is not involved in the training process of the model. TrUNet compared to the U-Net network, Dice improved from 80.44% to 87.83%, an increase of 7.39%. IoU improved from 71.98% to 80.27%, an increase of 8.29%. Accuracy improved from 92.10% to 94.58%, an increase of 2.48%. Recall improved from 78.04% to 88.88%, an increase of 10.84%. Precision decreases from 91.45% to 90.88%, a decrease of 0.57%. HD decreases from 7.94 to 4.21, a decrease of 3.73, with smaller HD values indicating better segmentation, reflecting better boundary alignment and overall performance.

The results clearly demonstrate that the combination of CNN and Transformer in TrUNet can give full play to the advantages of both, accurately predict the location and boundary of skin lesions, and effectively solve the problems of over-segmentation and under-segmentation. This structure is advantageous for skin lesion segmentation to achieve accurate segmentation. The visualization of the segmentation results of different algorithms on the ISIC2017 dataset are shown in FIGURE 6.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3463713

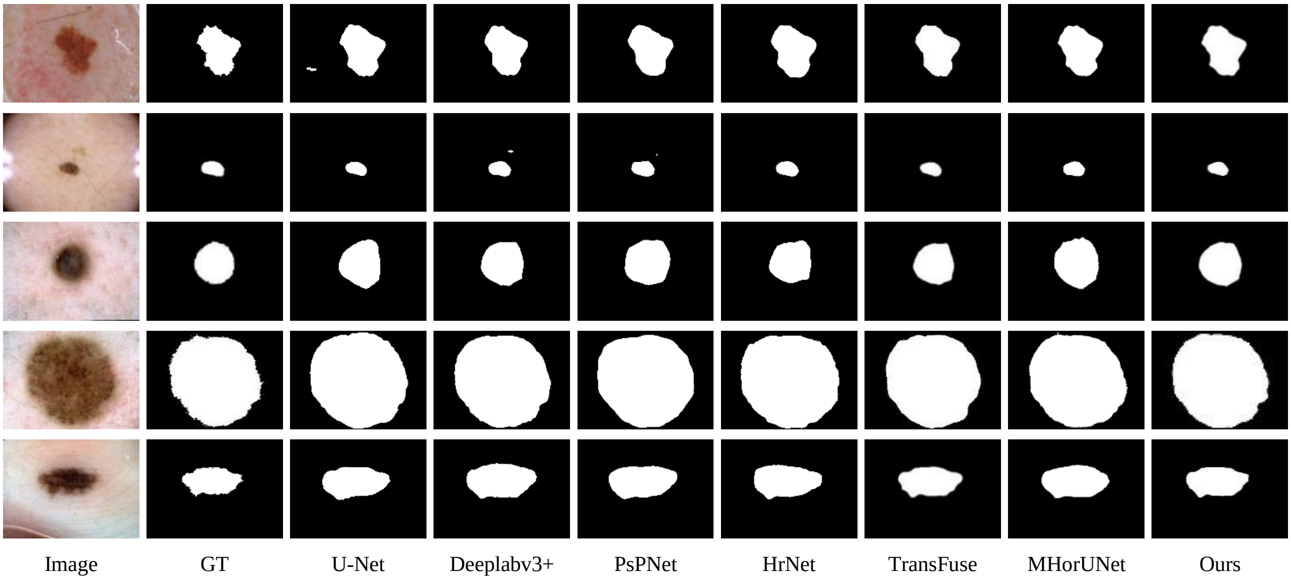

FIGURE 6. **Visual results of different methods on the ISIC2017 dataset.**

TABLE II
TABLE IIISEGMENTATION PERFORMANCE OF DIFFERENT METHODS ON THE ISIC2017 DATASET

Method	Dice↑	IoU↑	Accuracy↑	Recall↑	Precision↑	HD↓
U-Net[3]	80.44%	71.98%	92.10%	78.04%	91.45%	7.94
Deeplabv3+[11]	82.64%	74.29%	92.98%	80.15%	93.16%	7.53
PsPNet[10]	78.21%	69.12%	92.08%	75.33%	92.36%	7.72
HrNet[13]	81.46%	72.77%	92.71%	80.51%	90.57%	8.09
TransFuse[18]	86.13%	78.05%	93.95%	83.72%	**93.61%**	4.28
MHorUNet[17]	82.67%	70.46%	92.47%	76.13%	90.44%	**3.77**
Ours	**87.83%**	**80.27%**	**94.58%**	**88.88%**	90.88%	4.21

2) RESULTS ON THE PH2 DATASET

To verify the generalization ability of the TrUNet model, the ISIC2017 dataset was used for training, while the PH2 dataset served as an independent test set. The training weights from ISIC2017 were first loaded. Subsequently, TrUNet's test results on PH2 were compared with those of mainstream methods to evaluate its performance, and the results are shown in TABLE III. The Dice of TrUNet is 90.97%, the IoU is 83.98%, the Accuracy is 94.89%, the Recall is 98.81%, the Precision is 84.97%, and the HD is 4.48. While TrUNet did not achieve the highest Precision and HD, it still demonstrated the best performance across the remaining four evaluation metrics.

TrUNet compared to the U-Net network, Dice improved from 85.77% to 90.97%, an increase of 5.20%. IoU improved from 76.86% to 83.98%, an increase of 7.12%. Accuracy improved from 91.20% to 94.89%, an increase of 3.69%. Recall improved from 87.16% to 98.81%, an increase of 11.65%. Precision decreases from 87.47% to 84.97%, a decrease of 2.50%. HD decreases from 10.77 to 4.48, a decrease of 6.29. The visualization of the segmentation results of different algorithms on the PH2 dataset are shown in FIGURE 7.

TABLE III
SEGMENTATION PERFORMANCE OF DIFFERENT METHODS ON THE PH2 DATASET

Method	Dice↑	IoU↑	Accuracy↑	Recall↑	Precision↑	HD↓
U-Net[3]	85.77%	76.86%	91.20%	87.16%	**87.47%**	10.77
Deeplabv3+[11]	85.91%	76.32%	91.91%	92.60%	82.70%	11.18
PsPNet[10]	89.99%	82.54%	93.81%	96.74%	85.43%	9.31
HrNet[13]	89.91%	82.36%	93.50%	96.76%	85.25%	9.22
TransFuse[18]	90.83%	83.96%	94.15%	97.92%	85.98%	4.67
MHorUNet[17]	88.61%	77.74%	92.11%	89.32%	86.71%	**4.19**
Ours	**90.97%**	**83.98%**	**94.89%**	**98.81%**	84.97%	4.48

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3463713

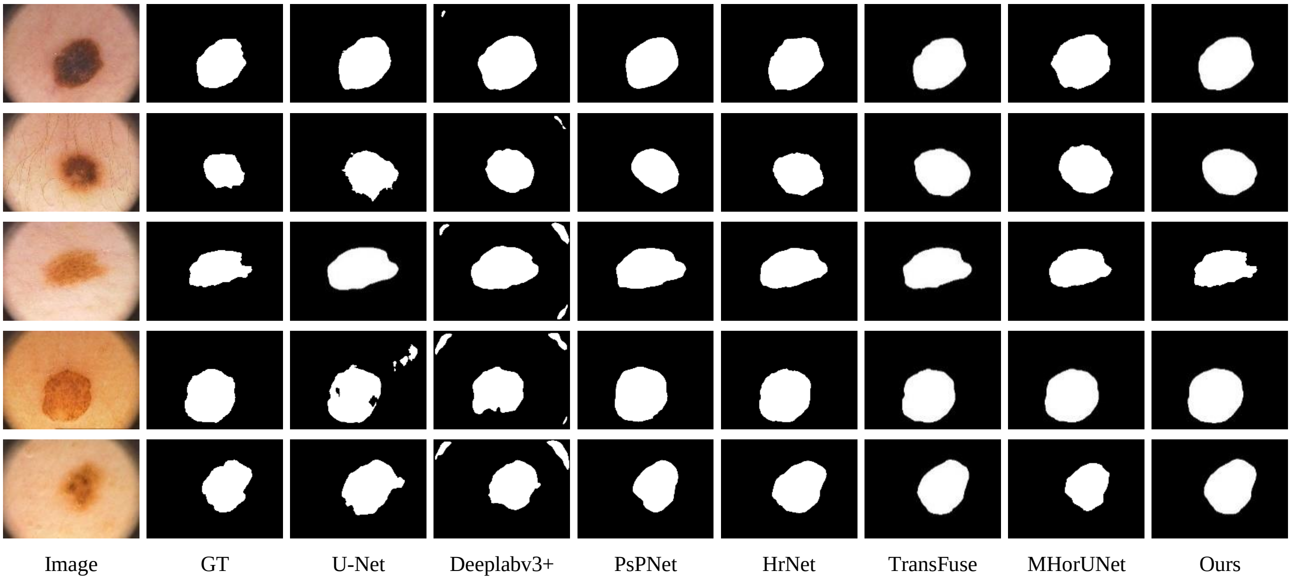

FIGURE 7. **Visual results of different methods on the PH2 dataset.**

3) RESULTS ON THE ISIC2018 DATASET

The experimental results of comparing the TrUNet model proposed in this paper with the current mainstream methods on the ISIC2018 dataset are shown in TABLE IV. The Dice of TrUNet is 90.61%, the IoU is 84.25%, the Accuracy is 94.74%, the Recall is 94.27%, the Precision is 89.54% and the HD is 4.62. Although TrUNet is not the best on Recall and HD, it produces the best performance on the other four metrics.

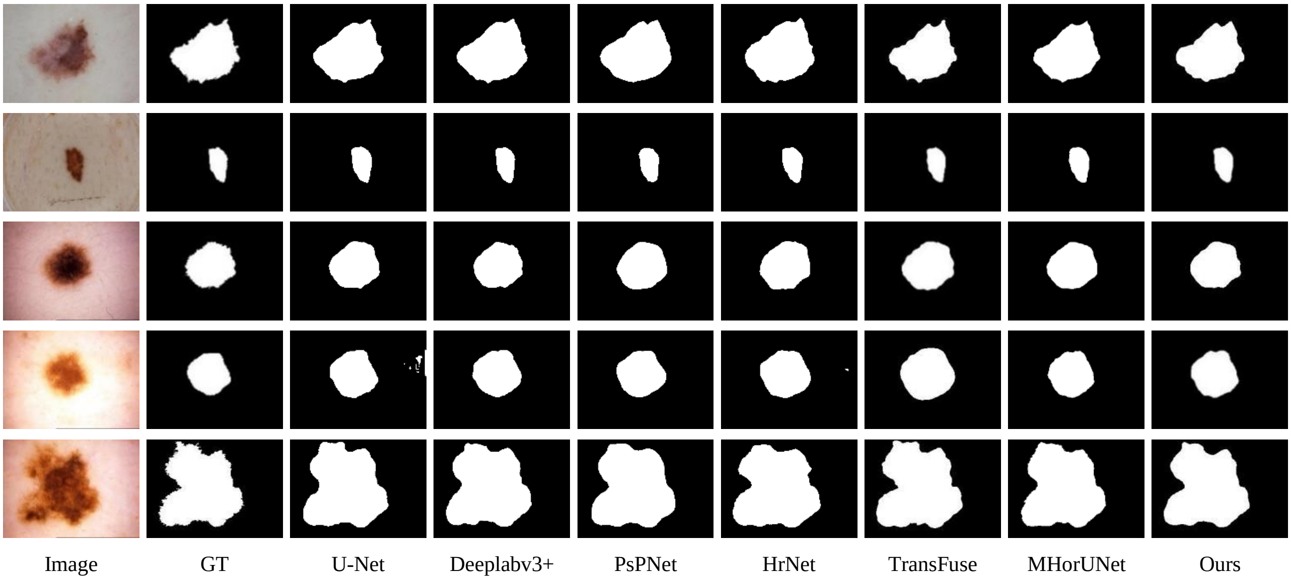

FIGURE 8. **Visual results of different methods on the ISIC2018 dataset**

TABLE IV
SEGMENTATION PERFORMANCE OF DIFFERENT METHODS ON THE ISIC2018 DATASET

Method	Dice ↑	IoU ↑	Accuracy ↑	Recall ↑	Precision ↑	HD ↓
U-Net[3]	82.89%	73.47%	90.76%	93.46%	78.45%	15.66
DeepLabv3+[11]	87.00%	78.90%	92.81%	92.37%	85.43%	15.58
PsPNet[10]	87.49%	79.55%	93.23%	92.07%	86.36%	15.14
HrNet[13]	86.43%	77.95%	92.30%	**94.99%**	82.05%	16.30
TransFuse[18]	88.67%	81.35%	94.00%	94.49%	86.36%	4.75
MHorUNet[17]	85.71%	74.99%	92.31%	82.36%	89.33%	**4.03**
Ours	**90.61%**	**84.25%**	**94.74%**	94.27%	**89.54%**	4.62

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3463713

TrUNet compared to the U-Net network, Dice improved from 82.89% to 90.61%, an increase of 7.72%. IoU improved from 73.47% to 84.25%, an increase of 10.78%. Accuracy improved from 90.76% to 94.74%, an increase of 3.98%. Recall improved from 93.46% to 94.27%, an increase of 0.81%. Precision improved from 78.45% to 89.54%, an increase of 11.09%. HD decreases from 15.66 to 4.62, a decrease of 11.04. The HD metric results show that TrUNet is more sensitive to boundary features and can accurately depict the target region. In order to demonstrate more intuitively the difference between TrUNet and the current popular methods, the visualization of the segmentation results of different algorithms on the ISIC2018 dataset are shown in FIGURE 8.

As can be seen from the figure, for smaller target regions, the prediction results of TrUNet are still optimal, and it can capture the remote dependencies well and reduce the loss of feature information. Obviously, the method proposed in this paper demonstrates the capability to segment the lesion edge region with remarkable accuracy. Even for minuscule lesions, it can achieve precise lesion localization and segmentation of lesion boundaries, showcasing its robust performance.

4) ABLATION STUDY

TrUNet is a dual-branch encoder-decoder network composed of the MFA module, TrFusion module, Res2Net branch, and Transformer branch. To verify the validity of individual modules in the TrUNet model, individual modules were selectively removed for ablation experiments, where PH2 was not used as an additional test data set. The results of the experiments are shown in TABLE V, and the visualization of the segmentation results are shown in FIGURE 9.

In the ablation experiments, the segmentation performance is evaluated by systematically removing each module from the TrUNet model and examining the resultant impact. By comparing the results of these ablation experiments, the contribution of each module can be accurately assessed, thereby leading to a deeper understanding of the overall performance of the TrUNet model and the roles of its individual components. In this paper, Transformer and Res2Net are chosen as the encoder branches of the baseline network to jointly construct the baseline network.

In summary, the TrUNet model designed in this paper is superior to other methods for the following reasons: (1) Combining CNN and Transformer for feature extraction provides richer information than when only a single network is used for feature extraction. (2) The preprocessing operation effectively solves the problems of hair occlusion, low contrast and boundary blurring in skin lesions, clarifies the segmentation target, and improves the segmentation accuracy. (3) Designing the MFA module to fuse different scale features extracted from the same branch at different stages with each other to complement spatial and positional information to achieve accurate grasp of segmentation edges. (4) Designing the TrFusion module to selectively fuse features extracted from different branches at the same stage to enhance important information and suppress irrelevant details. This fusion is not a simple addition or splicing, but reduces the effect of redundant information brought by two-branch fusion, and improves the robustness and generalization ability of the model.

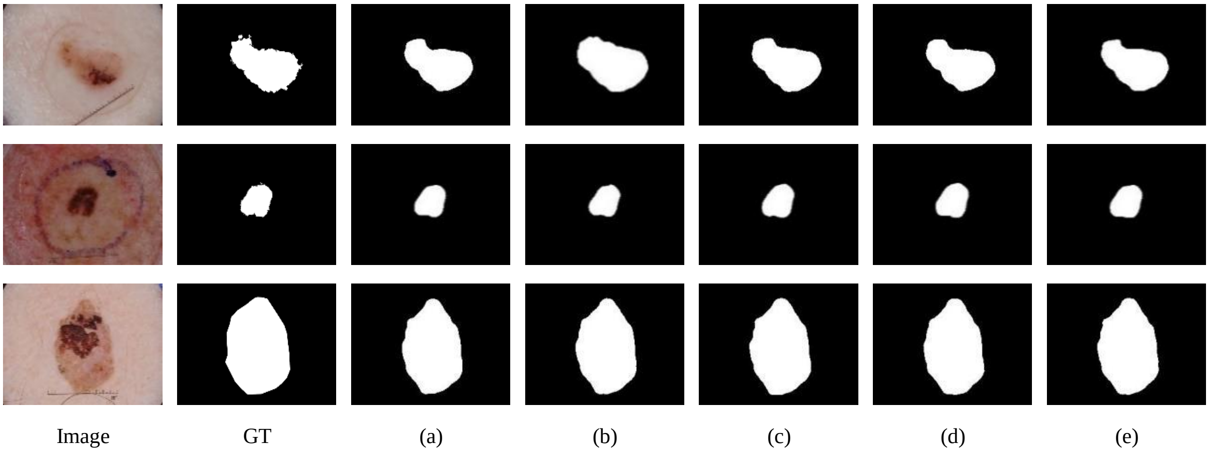

FIGURE 9. Visualization of ablation study of TrUNet on skin lesion segmentation in ISIC2017 dataset.

TABLE V
ABLATION STUDY OF TRUNET FOR SKIN LESION SEGMENTATION ON THE ISIC2017 DATASET.

	Method	Dice↑	IoU↑	Accuracy↑	Recall↑	Precision↑	HD↓
(a)	baseline+Pre-process	85.34%	77.31%	94.05%	81.67%	94.05%	4.11
(b)	baseline+Pre-process+TrFusion	86.04%	78.15%	94.19%	83.77%	94.19%	4.14
(c)	baseline+Pre-process+MFA	86.58%	78.69%	94.34%	84.42%	94.34%	4.25
(d)	baseline +TrFusion+MFA	86.08%	78.10%	94.04%	84.53%	94.04%	4.27
(e)	baseline+Pre-process+TrFusion+MFA	**86.78%**	**79.03%**	**94.49%**	**85.55%**	**94.49%**	**4.1**

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3463713

V. CONCLUSION

In this study, an overview of the application of CV and image processing techniques in skin lesion segmentation is initially presented, accompanied by an analysis of prevalent methods. Subsequently, the skin lesion segmentation network based on deep learning was introduced in detail. To address issues such as hair occlusion and low contrast in skin lesion images, preprocessing methods are proposed to improve hair occlusion and enhance the contrast between lesion and surrounding areas. To tackle the problem of insufficient feature interaction, an MFA module is introduced to supplement positional and spatial information, aiming to enhance the model's representation capability. Addressing the inadequate fusion of features from the two branches of the encoder, a TrFusion module is proposed to selectively fuse feature information extracted by the encoder branches and utilize attention mechanisms to suppress irrelevant details, thus improving network performance. Experimental results demonstrate that the proposed TrUNet network exhibits competitive segmentation performance on multiple datasets. Future research will focus on further enhancing the performance of skin lesion segmentation to achieve more precise segmentation results.

REFERENCES

[1] Chatterjee S, Dey D, Munshi S. Integration of morphological preprocessing and fractal based feature extraction with recursive feature elimination for skin lesion types classification[J]. Computer methods and programs in biomedicine, 2019, 178: 201-218.

[2] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[3] Ronneberger O, Fischer P, Brox T: U-Net: Convolutional networks for biomedical image segmentation[C]. Proceedings of the Medical Image Computing and Computer-Assisted Intervention, 2015: 234-241.

[4] 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, September 20, 2018, Proceedings 4. Springer International Publishing, 2018: 3-11.

[5] Qin X, Zhang Z, Huang C, et al. U2-Net: Going deeper with nested U-structure for salient object detection[J]. Pattern recognition, 2020, 106: 107404.

[6] Huang H, Lin L, Tong R, et al. Unet 3+: A full-scale connected unet for medical image segmentation[C]//ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2020: 1055-1059.

[7] Xiao X, Lian S, Luo Z, et al. Weighted res-unet for high-quality retina vessel segmentation[C]//2018 9th international conference on information technology in medicine and education (ITME). IEEE, 2018: 327-331.

[8] Li X, Chen H, Qi X, et al. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes[J]. IEEE transactions on medical imaging, 2018, 37(12): 2663-2674.

[9] Lin W, Chu J, Leng L, et al. Feature disentanglement in one-stage object detection[J]. Pattern Recognition, 2024, 145: 109878.

[10] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.

[11] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[12] Gu Z, Cheng J, Fu H, et al. Ce-net: Context encoder network for 2d medical image segmentation[J]. IEEE transactions on medical imaging, 2019, 38(10): 2281-2292.

[13] Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(10): 3349-3364.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]. Proceedings of the Advanes in Neural Information Processing Systems, 2017, 30: 6000-6010.

[15] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.

[16] Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 205-218.

[17] Wu R, Liang P, Huang X, et al. MHorUNet: High-order spatial interaction UNet for skin lesion segmentation[J]. Biomedical Signal Processing and Control, 2024, 88: 105517.

[18] Zhang Y, Liu H, Hu Q. Transfuse: Fusing transformers and cnns for medical image segmentation[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer International Publishing, 2021: 14-24.

[19] Gao S H, Cheng M M, Zhao K, et al. Res2net: A new multi-scale backbone architecture[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(2): 652-662.

[20] 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, September 20, 2018, Proceedings 4. Springer International Publishing, 2018: 3-11.

[21] Jha D, Smedsrud P H, Riegler M A, et al. Resunet++: An advanced architecture for medical image segmentation[C]//2019 IEEE international symposium on multimedia (ISM). IEEE, 2019: 225-2255.

[22] Liu X, Song L, Liu S, et al. A review of deep-learning-based medical image segmentation methods[J]. Sustainability, 2021, 13(3): 1224.

[23] Yuan Y, Chao M, Lo Y C. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance[J]. IEEE transactions on medical imaging, 2017, 36(9): 1876-1886.

[24] He X, Yu Z, Wang T, et al. Skin lesion segmentation via deep RefineNet[C]. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2017.

[25] Bi L, Feng D, Fulham M, et al. Improving skin lesion segmentation via stacked adversarial learning[C]//2019 IEEE 16Th International symposium on biomedical imaging (ISBI 2019). IEEE, 2019: 1100-1103.

[26] Huang L, Zhao Y, Yang T. Skin lesion segmentation using object scale-oriented fully convolutional neural networks[J]. Signal, Image and Video Processing, 2019, 13: 431-438.

[27] Berseth M. ISIC 2017 - Skin Lesion Analysis Towards Melanoma Detection[J]. Computer Vision and Pattern Recognition, 2017, 125(3): 58-65.

[28] Tang Y, Yang F, Yuan S, et al. A multi-stage framework with context information fusion structure for skin lesion segmentation[C]//2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). 2019: 1407-1410.

[29] Schlemper J, Oktay O, Schaap M, et al. Attention gated networks: Learning to leverage salient regions in medical images[J]. Medical image analysis, 2019, 53: 197-207.

[30] Alom M Z, Yakopcic C, Hasan M, et al. Recurrent residual U-Net for medical image segmentation[J]. Journal of medical imaging, 2019, 6(1): 014006-014006.

[31] 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.

[32] Song P, Li J, Fan H, et al. TGDAUNet: Transformer and GCNN based dual-branch attention UNet for medical image segmentation[J]. Computers in Biology and Medicine, 2023, 167: 107583.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3463713

[33] Chen Y, Wang T, Tang H, et al. CoTrFuse: a novel framework by fusing CNN and transformer for medical image segmentation[J]. Physics in Medicine & Biology, 2023, 68(17): 175027.

[34] 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. 2021: 10012-10022.

[35] Chen B, Liu Y, Zhang Z, et al. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2023.

[36] Lin A, Chen B, Xu J, et al. Ds-transunet: Dual swin transformer u-net for medical image segmentation[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-15.

[37] Codella N C F, Gutman D, Celebi M E, 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)[C]//2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE, 2018: 168-172.

[38] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, Allan Halpern: "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)", 2018; https://arxiv.org/abs/1902.03368

[39] Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi:10.1038/sdata.2018.161(2018).

[40] Mendonça T, Ferreira P M, Marques J S, et al. PH 2-A dermoscopic image database for research and benchmarking[C]//2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, 2013: 5437-5440.

Wei Chen received the B.S. and M.S. degrees from Zhejiang University, and the Ph.D. degree from the University of Chinese Academy of Sciences. He is currently an Associate Professor with the School of Communication and Information Engineering, Xi’an University of Science and Technology. His current research interests include the areas of computer vision, artificial intelligence, image processing, and optoelectric detection.

Qian Mu received a B.S degree from Jishou University. She is currently pursuing a Master's degree in Electronic Information at Xi'an University of Science and Technology. Her research focuses on computer vision, artificial intelligence, and image processing.

Jie Qi received the B.S., M. S. and Ph.D degrees from West China Medical Center, Sichuan University. She is currently a Chief Physician at the Orthopedic Department, Shaanxi Provincial People's Hospital. Her current research interests include the areas of anatomical feature recognition, image segmentation and image analysis.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4