Biomedical Signal Processing and Control 100 (2025) 107090

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

DSU-Net: Dual-Stage U-Net based on CNN and Transformer for skin lesion segmentation

Longwei Zhong [a], Tiansong Li [a,*], Meng Cui [a], Shaoguo Cui [a], Hongkui Wang [b], Li Yu [c]

[a] College of Computer and Information Science, Chongqing Normal University, Chongqing, China
[b] School of Telecommunications Engineering, Hangzhou Dianzi University, Hangzhou, China
[c] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China

ARTICLE INFO

Dataset link: https://github.com/ZhongLongwei/DSU-Net

Keywords:
Skin lesion segmentation
CNNs
Transformer
Dual-stage U-Net
Two-stage balanced loss function
Multi-feature fusion

ABSTRACT

Precise delineation of skin lesions from dermoscopy pictures is crucial for enhancing the quantitative analysis of melanoma. However, this remains a difficult endeavor due to inherent characteristics such as large variability in lesion size, form, and fuzzy boundaries. In recent years, CNNs and Transformers have indicated notable benefits in the area of skin lesion segmentation. Hence, we first propose the DSU-Net segmentation network, which is inspired by the manual segmentation process. Through the coordination mechanism of the two segmentation sub-networks, the simulation of a process occurs where the lesion area is initially coarsely identified and then meticulously delineated. Then, we propose a two-stage balanced loss function to better simulate the manual segmentation process by adaptively controlling the loss weight. Further, we introduce a multi-feature fusion module, which combines various feature extraction modules to extract richer feature information, refine the lesion area, and obtain accurate segmentation boundaries. Finally, we conducted extensive experiments on the ISIC2017, ISIC2018, and PH2 datasets to assess and validate the efficacy of the DSU-Net by comparing it to the most advanced approaches currently available. The codes are available at https://github.com/ZhongLongwei/DSU-Net.

1. Introduction

Melanoma, an aggressive tumor of melanocytes, is the most lethal skin cancer. Depending on the most recent cancer data published by the American Cancer Society [1] project, the United States is expected to experience an estimated 104,930 new cases of skin cancer in 2023. Melanoma patients comprise 97,610 of these cases. Skin lesions are segmented from skin pictures, which is important in the diagnosis and treatment of skin cancer. Current diagnostic methods mainly rely on manual inspection. However, the manual examination is a tedious task that can be subjective, non-repeatable, and heavily reliant on the physician's experience. Thus, it is significant to achieve automated segmentation of skin diseases in computer-aided diagnostic (CAD) systems.

However, due to the influence of several factors, skin lesion segmentation is a complicated and challenging endeavor. A typical case is shown in Fig. 1. First and foremost, the skin lesion region differs in color, shape, and size. Then, it is vulnerable to influence from other elements, such as blood vessels, hair, and skin texture. Finally, the contrasting effect between the skin lesion and surrounding skin is minuscule, and its borders are vague. To address these issues, numerous classic skin lesion segmentation methodologies have been provided to employ image processing technology, such as the threshold segmentation method [2–5], clustering method [6,7], and region merging method [8], among others. Nevertheless, these traditional methods typically only use shallow image features and struggle to handle lesion images in complex situations.

In recent years, with the rise of deep learning, convolutional neural networks (CNNs) have emerged as a new choice. CNNs were able to extract deep features from images and achieve excellent model performance without manual intervention. In particular, the creation of U-Net [9] promoted the emergence of a series of network models based on it, such as UNet++ [10], ResUNet [11], DenseNet [12], AttU-Net [13], etc. To address the challenges of skin lesion segmentation tasks, Yu et al. [14] presented the fully convolutional residual network (FCRN) to extract representative skin lesion segmentation characteristics using residual connections and multi-scale contextual information. Tang et al. [15] employed detachable convolutional blocks and U-Net structure to augment the discriminatory presentation abilities of the fully convolutional network (FCN) at the pixel level. They also presented a technique based on random weighted averaging to boost

* Corresponding author.
E-mail address: tiansongli@cqnu.edu.cn (T. Li).

https://doi.org/10.1016/j.bspc.2024.107090
Received 26 June 2024; Received in revised form 7 October 2024; Accepted 15 October 2024
Available online 29 October 2024

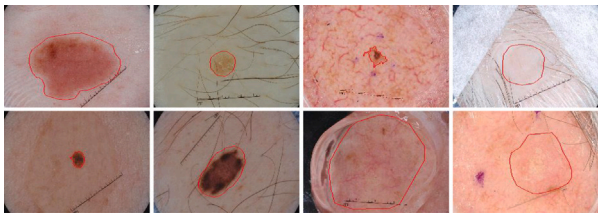

Fig. 1. Some challenging common lesion samples from publicly available dermoscopic pictures. (a) Different shapes and sizes. (b) Interference by hair. (c) Interference by blood vessels. (d) Low contrast. The red line in the picture is the real label.

the model's generalization ability. However, due to the inherent limitations of convolution, CNN only has a local receptive field and cannot effectively model long-term dependencies, which limits its ability to further refine segmentation in skin lesion segmentation tasks. Within the domain of medical picture semantic segmentation, the long-term dependencies involved in pixel allocation are of enormous significance for accurately defining boundary pixels, especially skin lesion boundary pixels. Therefore, by enhancing the global contextual information of feature maps and learning the long-range dependencies between pixels in medical images, segmentation performance can be effectively improved, enabling precise localization of lesion areas and boundaries.

Simultaneously, the proposal of Transformer [16] and the efficient use of Vision Transformer (ViT) [17] in image classification tasks offered novel approaches to address this issue. Many Transformer-based networks for image segmentation tasks have been proposed one after another, such as Swin-Unet [18], nnFormer [19], MISSfomer [20], etc. Although Transformer performed well in modeling contextual dependencies, it is relatively insufficient in extracting local information. Therefore, hybrid network structures based on CNN and Transformer have become a new research hotspot. For instance, Wu et al. [21] introduced FAT-Net, which efficiently combined CNN and Transformer to acquire local and global context information for skin lesion segmentation using a dual-branch structure. He et al. [22] integrated multi-scale channel attention features, local features captured by CNNs, and global features captured by Transformer into just one module, significantly enhancing the feature representation capabilities. Although CNN and Transformer-based methods have made significant progress in skin lesion segmentation tasks, they still exhibit clear limitations in certain scenarios. For example, when handling blurred edges and substantial background interference, these methods often struggle to accurately identify lesion boundaries. Inspired by the manual segmentation process, we propose a two-step strategy to refine the segmentation task and more effectively address these challenges.

In this paper, we present a dual-stage U-Net (DSU-Net) based on CNNs and Transformers. This network draws on the manual segmentation process, which simulates the process of first coarse segmentation and then refinement through the collaborative mechanism between the two-stage segmentation sub-networks. During the first stage, we employ a Swin Transformer-based rough segmentation network to simulate the process of identifying rough areas of lesions. In the second stage, a refined segmentation network containing more feature information is used to simulate the process of refining the lesion area. This design can not only achieve efficient lesion segmentation but also remove some interference factors in the first stage and further reduce the area to be segmented, preparing for the next stage of segmentation. Meanwhile, to strengthen the synergy between segmentation networks, we present a two-stage balanced loss function, which allows the network to learn the corresponding target area adaptively by adjusting the weight ratio of the loss. In addition, inspired by the H2Former [22], we devised a Multi-Feature Fusion Module (MFFM) that can improve the proposed DSU-Net ability to detect and understand lesions of different scales by fusing global and local multi-scale features, thereby assisting in accurate boundary localization. Finally, we perform extensive experiments on the ISIC2017, ISIC2018, and PH2 datasets to validate the efficacy of the proposed DSU-Net by comparing it with the most advanced approaches currently available and conducting ablation studies. The following is a brief overview of our contributions:

(1) We present a dual-stage U-Net (DSU-Net) based on CNNs and Transformers. This network utilizes a collaborative mechanism between the two-stage segmentation sub-networks, first performing coarse segmentation and then refining the lesion area. This approach effectively reduces interference, leading to more precise lesion segmentation.
(2) We introduce a two-stage balanced loss function to enhance collaboration between the segmentation sub-networks, allowing adaptive learning of target areas through loss weight adjustment.
(3) We present a multi-feature fusion module (MFFM) that enhances the network's ability to detect and understand lesions at different scales by extracting and combining multi-scale global and local features, thereby improving the segmentation accuracy of lesion boundaries.

2. Related work

2.1. CNNs-based method

Recently, CNN-based methods have been extensively employed in medical image segmentation, like FCN [23], UNet [9], DeepLabv3 [24], SegNet [25], UNet++ [10], CE-Net [26], MaMfi-Net [27], MSS-UNet [28], etc. Skin lesion segmentation is a specialized domain within medical image segmentation that targets accurately distinguishing skin lesion areas at the pixel level. To achieve effective skin lesion segmentation, CNN-based methods have been continuously presented. For example, Yuan et al. [29] designed a new loss function utilizing the Jaccard distance to solve the issue of unbalanced foreground and background pixels in image segmentation when employing cross-entropy loss. Xie et al. [30] devised a mutually guided deep convolutional neural network (MB-DCNN) to enhance the veracity of segmenting and classifying skin lesions together. Bagheri et al. [31] proposed a two-stage automatic segmentation method for skin lesions, combining the CNN-based Retina-Deeplab and Mask R-CNN detection segmentation structures, and improving segmentation accuracy through a geodesic-based/graph-based combination strategy.

Although these methods significantly enhance the segmentation capability of the model, they neglect the utilization of global information. To obtain more valuable information, researchers have introduced attention mechanisms to extract richer features. For example, Dai et al. [32] presented a new multi-scale residual encoding and decoding network (MsRED) for skin lesion segmentation while designing a multi-resolution and multi-channel feature fusion module (M2F2) to improve the model's representation learning capacity. Hu et al. [33] presented the attention synergy network (AS-Net), which efficiently combined channel and spatial information using a coordination module to increase the model's recognition capacity. Wu et al. [34] introduced the adaptive dual attention module (ADAM), which combines two complimentary global context modules to effectively utilize global information from different angles and enhance the veracity of skin lesion segmentation. Ruan et al. [35] presented MALUNet, which extracts feature information through multiple attention mechanisms that decrease model parameters while retaining performance. Arora et al. [36] utilized attention gates (AG) to differentiate between high-dimensional information and low-level irrelevant areas of the background in the input image, effectively improving segmentation performance. Sun et al. [37] proposed the multi-scale context attention network (MSCA-Net), which combined multi-scale context and global-local channel spatial attention to address the issue of spatial and context information loss arising from image resolution reduction.

2.2. Transformer based methods

In recent years, the Transformer [16], originally designed for natural language processing, has been successfully extended to other domains, gaining substantial attention in fields like image classification, detecting objects, and semantic segmentation. Vision Transformer (ViT) [17] is the first Transformer model to achieve classification duties in computer vision. ViT trains by dividing images across non-overlapping patches, demonstrating performance comparable to or even surpassing convolutional neural networks (CNNs). Transformer-based segmentation models have also been proposed as a result of the exceptional performance of Transformers in computer vision, such as [18,38–41]. Among them, Cao et al. [18] presented Swin-Unet, which replaces the encoder and decoder parts of U-Net with the Swin Transformer [42], thereby achieving superior segmentation performance. He et al. [40] proposed the Fully Transformer Network (FTN), which leverages the hierarchical computational characteristics of the Spatial Pyramid Transformer (SPT) to significantly enhance the performance of skin lesion segmentation. Liu et al. [41] proposed CSWin-UNet, which integrates the CSWin self-attention mechanism into UNet and introduces a novel decoder, significantly improving the computational efficiency and segmentation accuracy in medical image segmentation.

2.3. Hybrid network based CNN and transformer

Although Transformers have shown excellent performance in the segmentation domain, they are relatively weak at extracting local information, which sometimes leads to poor segmentation performance. Therefore, some researchers have proposed to integrate CNNs and Transformers [21,22,43–47]. Among them, Wang et al. [43] proposed a Boundary-Aware Transformer (BAT) that integrates CNN with Transformer networks. They introduced a Boundary Attention Gate (BAG) within the Transformer to enhance the recognition of skin lesion boundaries, thereby improving segmentation performance. Cao et al. [44] proposed ICL-Net, which enhances feature representation capability by complementarily leveraging semantic correlations between pixels through the pyramid transformer pixel inter-relationship correlation (PTIC) module and the local neighborhood metric learning (LNML) module. Huang et al. [47] presented ADF-Net, an innovative multi-stage dual-stream hybrid network that integrates coarse-grained and fine-grained features through an adaptive feature fusion module and a focal attention decoder, achieving efficient and accurate automatic skin lesion segmentation. Although the above methods have achieved excellent results, they all perform segmentation directly instead of adopting a two-stage segmentation approach from coarse to fine. Based on this, we present a two-stage hybrid transformer network. The network simulates the manual segmentation process and first uses Swin Transformer to preliminarily segment the rough lesion area. Secondly, in the process of further refining the boundary, CNN and self-attention mechanisms are utilized for acquiring channel, local, and global context information to enhance feature representation ability. Simultaneously, multi-scale global and local features are merged to increase the model's capacity to recognize and understand lesions of varying sizes, resulting in precise lesion boundary segmentation.

3. Methodology

3.1. Overview

The proposed model architecture is shown in Fig. 2, which is comprised of two segmentation sub-networks: a one-stage coarse segmentation network based on Swin Transformer (RoughSwinSegNet) and a two-stage refined segmentation network (RefineSegNet). DSU-Net utilizes the skin image as input and first performs preliminary segmentation through the RoughSwinSegNet based on Swin Transformer to obtain a rough prediction of the lesion area. Then, the rough prediction result is converted into the corresponding prediction category and multiplied with the input image, pixel by pixel. Subsequently, the processed data is input into the RefineSegNet for further refined segmentation, ultimately generating an accurate prediction label for the lesion area. In the next sections, we will go into detail about these two segmentation sub-networks.

3.2. RoughSwinSegNet

The RoughSwinSegNet aims to perform a rough segmentation of the lesion area to remove most of the background area and reduce the impact on factors like blood vessels and hair, so that the second stage can focus more on boundary recognition. As shown in Fig. 2(a), it utilizes Swin Transformer [42] as the encoder. The primary reason is that Swin Transformer employs the sliding window mechanism to construct hierarchical features, making it suitable for handling irregular-shaped skin lesion segmentation. In addition, Transformer has global long-distance modeling capabilities, which can effectively encode high-level feature representations and accurately locate lesion areas. To reduce the model's parameters while retaining performance, we use 1×1 convolution and transposed convolution to build the decoder, and straight addition for skip connections. In the meantime, we designed a two-stage balanced loss function to guide the RoughSwinSegNet to generate rough prediction results. Among them, the (0, 1) conversion operation converts the rough prediction results into the corresponding prediction categories.

3.3. RefineSegNet

The RefineSegNet is mainly a process of further refining the segmentation boundary based on the RoughSwinSegNet segmentation result, which can effectively deal with complex segmentation boundary situations. As depicted in Fig. 2(b), the RefineSegNet encoder comprises DownSample and Multi-feature Fusion Module (MFFM), whereas the decoder is the same as RoughSwinSegNet. Among them, the MFFM module consists of four basic units: channel attention module (CAM), multi-scale feature extraction module (MSFE), local attention module (LAM), and global attention module (GAM). As depicted in Fig. 3. The MFFM module can enhance RefineSegNet's ability to detect and understand lesion areas of different scales by extracting global and local information and fusing multi-scale global and local features. This helps to solve the challenges of lesion areas with varying sizes and shapes, as well as fuzzy boundaries, thereby providing more accurate segmentation results. In addition, inspired by DANet [48], CAM and GAM modules are designed to extract channel and global features. Next, the core modules CAM, GAM, LAM, and MSFE in RefineSegNet are introduced.

3.3.1. Channel Attention Module (CAM)

The Channel Attention Module (CAM) utilizes the mechanism of self-attention to dynamically assign weights to the channel dimensions of the map of features. By learning the weight of each channel, it can highlight beneficial feature channels and suppress channels that are irrelevant to the final feature representation, thereby enhancing the model's feature representation capabilities. As seen in Fig. 4. The CAM module first reshapes the input feature $F_{ic} \in R^{C\times H\times W}$ into F_a and F_x, where C, H, and W denote channel, height, and width, respectively. Then, the transpose matrix of F_x and F_a is multiplied, and the end product is sent to the Softmax layer to model the relationship between channels, thereby allowing the global channel attention map $M_c \in R^{C\times C}$. Next, the feature F_{ic} is reshaped and multiplied by the channel attention map M_c. The resulting output is reshaped again and added to the feature F_{ic}, yielding the final output feature F_{oc}.

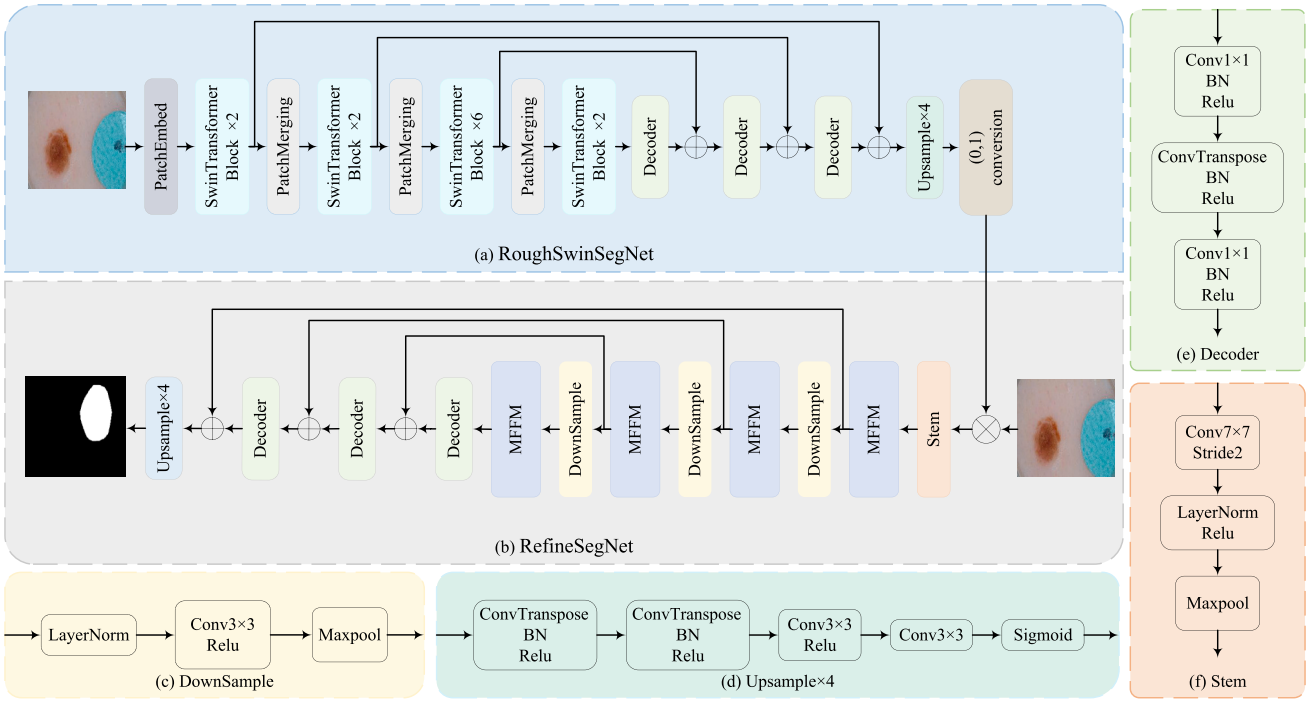

Fig. 2. The architecture of our proposed DSU-Net. (a) RoughSwinSegNet, (b) RefineSegNet, (c) RefineSegNet DownSample, (d) DSU-Net Upsample×4, (e) DSU-Net Decoder, (f) RefineSegNet Stem. This network initially performs preliminary segmentation using RoughSwinSegNet, which is based on the Swin Transformer, to obtain a rough prediction of the lesion area. The rough prediction is then converted into the corresponding prediction category and multiplied pixel by pixel with the input image. The processed data is subsequently fed into RefineSegNet for further refinement, ultimately producing an accurate prediction label for the lesion area.

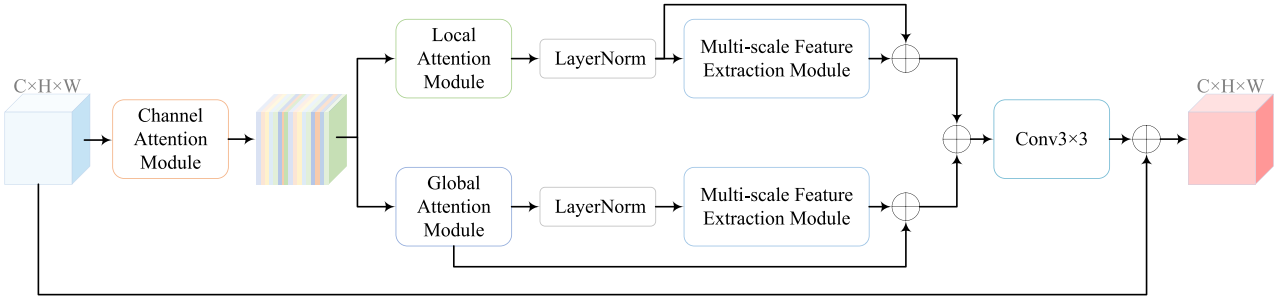

Fig. 3. The structures of Multi-feature Fusion Module (MFFM).

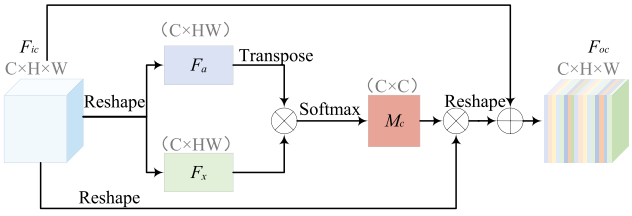

Fig. 4. The structures of Channel Attention Module (CAM).

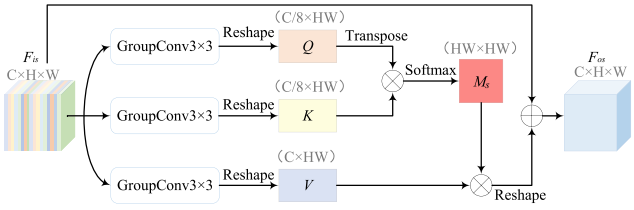

Fig. 5. The structures of Global Attention Module (GAM).

3.3.2. Global Attention Module (GAM)

GAM utilizes self-attention to extract the feature map's global context information. By weighting and summarizing the entire feature map, the network can have a wider range of perception capabilities when processing images, thereby improving the understanding of the global image structure and context. As shown in Fig. 5. The GAM module first generates Q, K, and V through three group convolutions of the input feature $F_{is} \in R^{C\times H\times W}$ and then performs Reshape. Among them, $Q \in R^{C/8\times HW}$, $K \in R^{C/8\times HW}$, and $V \in R^{C\times HW}$. Then, the transposed matrices of K and Q are multiplied, and the end product is sent to the Softmax layer to model the relationship between pixels in order to derive the global spatial attention map $M_s \in R^{HW\times HW}$. Next, V is multiplied by the spatial attention map M_s, and the resulting output is reshaped and added to the feature F_{is}, yielding the final output feature F_{os}.

3.3.3. Multi-scale Feature Extraction Module (MSFE)

The aim of MSFE is to strengthen the model's capacity to identify and comprehend lesion areas across different scales by extracting multi-scale information on local and global features. As depicted in Fig. 6. To

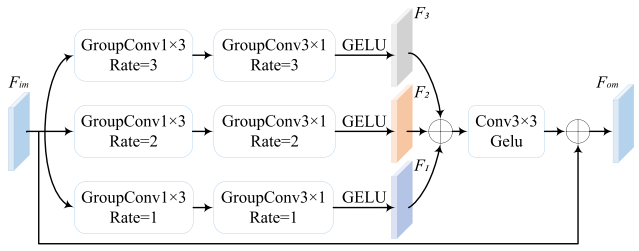

Fig. 6. The structures of Multi-scale Feature Extraction Module (MSFE).

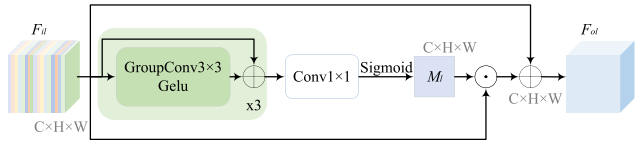

Fig. 7. The structures of Local Attention Module (LAM).

capture multi-scale feature information and expand the receptive field, we adopted a strategy that combines group convolution with dilated convolution. Additionally, by setting the convolution kernel to $(1,3)$ and $(3,1)$, the 2-dimensional convolution operation is converted into a linear complexity operation, which improves the model efficiency while decreasing model parameters. The input feature $F_{im} \in R^{C\times H\times W}$ first passes through three group convolutions with different dilated ratio and a kernel size of 1×3. It is then passed through three more group convolutions with different dilated ratio and a kernel size of 3×1 for feature extraction. After being activated by the GELU activation function, three features F_1, F_2, and F_3 are obtained. We sum the three obtained features and pass them through a 3×3 convolution followed by GELU activation for feature fusion. Finally, the fused features are added to F_{im}, resulting in the output feature F_{om}. The aforementioned procedure can be depicted as follows:

$$F_1 = GELU\left(gc2(gc1(F_{im}))\right) \tag{1}$$

$$F_2 = GELU\left(gc2'\left(gc1'(F_{im})\right)\right) \tag{2}$$

$$F_3 = GELU\left(gc2''\left(gc1''(F_{im})\right)\right) \tag{3}$$

$$F_{om} = F_{im} + GELU\left(Conv3(F_1 + F_2 + F_3)\right) \tag{4}$$

where $gc1$, $gc1'$, and $gc1''$ respectively represent group convolutions with dilated ratio of 1, 2, and 3 and a convolution kernel of $(1,3)$. $gc2$, $gc2'$, and $gc2''$ respectively represent group convolutions with dilated ratio of 1, 2, and 3 and a convolution kernel of $(3,1)$. $Conv3$ represents a 3×3 convolution.

3.3.4. Local attention module (LAM)

The role of LAM is to focus on specific local regions in the image to better capture local feature information. As depicted in Fig. 7. The LAM module initially passes the input feature $F_{il} \in R^{C\times H\times W}$ to the 3×3 Group Convolution layer for feature extraction. After being activated by the GELU, L is obtained, and the input before convolution is added to L through the skip connection. Repeat this process three times to obtain the feature F_l. Then, the local attention feature map $M_l \in R^{C\times H\times W}$ is generated by passing F_l into the 1×1 convolution layer and normalizing it with the sigmoid function. Lastly, the final output feature F_{ol} is generated by multiplying the local attention map M_l by the feature F_{il} through element-wise multiplication and adding it to the feature F_{il}. The aforementioned procedure can be depicted as follows:

$$F_{l''} = GELU\left(g(F_{il})\right) + F_{il} \tag{5}$$

$$F_{l'} = GELU\left(g(F_{l''})\right) + F_{l''} \tag{6}$$

$$F_l = GELU\left(g(F_{l'})\right) + F_{l'} \tag{7}$$

$$M_l = sigmoid(Conv1(F_l)) \tag{8}$$

$$F_{ol} = M_l \odot F_{il} + F_{il} \tag{9}$$

where g represents Group Convolution, $\odot$ represents element-wise multiplication, and $Conv1$ represents 1×1 convolution.

3.4. Two-stage balanced loss function

We employ two distinct loss functions to train the network. The initial function is the Dice loss function. Dice loss aims to assess how similar the predicted boundaries or regions are to the true labels. Optimizing Dice loss can help the model capture the border and shape of the target more accurately. The second loss function is Cross-Entropy. The aim of cross-entropy loss is to gauge the discrepancy between the probability distribution predicted by the model for each pixel and the actual label. The formula is as specified:

$$L_{dice} = 1 - 2 \times \frac{\sum_i P_i G_i}{\sum_i P_i + \sum_i G_i} \tag{10}$$

$$L_{bce} = -\sum_i \left[\left(1 - G_i\right) ln\left(1 - P_i\right) + G_i \ln\left(P_i\right)\right] \tag{11}$$

where i is an index for each pixel in the probability distribution map. P_i represents the probability that the ith pixel is part of the segmented region, whereas G_i denotes the actual label of the ith pixel.

However, from a large number of preliminary experiments, it can be found that simply applying the same loss function to the two segmentation sub-networks does not necessarily produce satisfactory results and may even lead to imprecise segmentation results. Therefore, to enhance the collaboration between two sub-networks and better simulate the segmentation process from coarse to fine, we present a two-stage balanced loss function. This function allows the network to learn the target area by adjusting the weight ratio of the loss in order to attain the optimal objective, as stated by:

$$L_1 = \varphi L_{bce} + (1 - \varphi) L_{dice} \tag{12}$$

$$L_2 = \mu L_{bce} + (1 - \mu) L_{dice} \tag{13}$$

$$L_{total} = \alpha L_1 + \beta L_2 \tag{14}$$

where L_1 denotes the loss of RefineSegNet, L_2 denotes the loss of RoughSwinSegNet, and φ and μ are assigned to 0.6 and 0.6, respectively. It has been determined that the optimal values for α and β are 2.0 and 1.6, following an extensive experimental evaluation. The specific process is as follows: First, the values of the parameters φ, μ, and α were fixed at 0.6, 0.6, and 2.0, respectively. Then, by adjusting the value of β, validation was conducted on the ISIC2017 dataset to determine the optimal loss coefficient. As shown in Fig. 8, when $\beta = 1.6$, the model achieved its best performance.

4. Experiments

4.1. Datasets

We gauged the effectiveness of our method on three extensively utilized public skin lesion segmentation benchmark datasets. These three datasets were contributed by the International Skin Imaging Collaboration (ISIC) and the Department of Dermatology at Pedro Hispano Hospital in Matosinhos, Portugal. They are the ISIC 2017 dataset [49], ISIC 2018 dataset [50], and PH2 dataset [51]. Data distribution for these datasets is as follows:

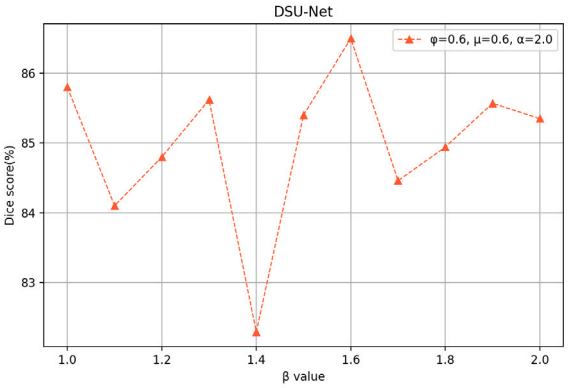

Fig. 8. Experimental results of DSU-Net with different loss function coefficients.

- **ISIC 2017.** The ISIC 2017 dataset has 2750 dermoscopic images that have been annotated. The images are categorized into three sets: training, validation, and testing. The training set comprises 2000 images, the validation set is comprised of up to 150 images, and the test set contains 600 images.
- **ISIC 2018.** The ISIC 2018 dataset consists of 3694 dermatology images. Out of them, 2594 images are assigned for training, 100 for validation, and the remaining 1000 for testing.
- **ISIC PH2.** The PH2 dataset consists of 200 dermoscopic images. The dataset is randomly divided into 140 images for training, 20 for validation, and 40 for testing in a ratio of 7:1:2.

4.2. Implementation details

We utilize a 12 GB NVIDIA GeForce RTX 3060 GPU to implement DSU-Net on PyTorch. In this study, the resolution of all images in the dataset is uniformly adjusted to 224 × 224. At the same time, to obtain better model initialization, the Swin Transformer [42] part of the one-stage coarse segmentation network uses the pre-trained model swin-tiny-patch4-window7-224. Among them, the channels of each layer of RoughSwinSegNet are [96,192,384,768], and the channels of each layer of RefineSegNet are [72,144,256,320].

We employ Adam with Weight Decay Fix (AdamW) as our foundational optimizer. The initial learning rate is defined as 0.0001, the betas parameter is specified as (0.9, 0.99), and the weight decay is provided as 0.00005. Simultaneously, a polynomial learning rate decay strategy is used to adaptively modify the learning rate, represented by the following expression:

$$LR_i = LR_{i-1} \times \left(1 - \frac{i}{Total\_epoch + 1}\right)^{\theta} \tag{15}$$

where i denotes the ith training cycle. The starting rate of learning is denoted as LR_0. The total number of training epochs is denoted as $Total\_epoch$. The learning rate of the ith epoch is denoted as LR_i, and θ is set to 0.9. The weights obtaining optimal performance on the validation set are chosen as the final test model after all networks have undergone training for 50 epochs. To augment the diversity of picture samples, we adopt the identical data enhancement method as [21]. The input image is vertically flipped along the x-axis or y-axis and rotated by an arbitrary angle between −15 and 15 degrees. Additionally, the brightness and contrast of the input image are randomly adjusted by a factor of −3% to 3%, and the hue, saturation, and value are also randomly altered within the same range of −3% to 3%.

4.3. Evaluation metrics

We employ five commonly utilized metrics to gauge the effectiveness of the method we propose, namely specificity (SP), sensitivity (SE), intersection and union (IoU), dice coefficient (Dice), and accuracy (ACC).

$$SE = \frac{TP}{TP + FN} \tag{16}$$

$$SP = \frac{TN}{TN + FP} \tag{17}$$

$$ACC = \frac{TP + FN}{TP + TN + FP + FN} \tag{18}$$

$$Dice = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{19}$$

$$IOU = \frac{TP}{TP + FP + FN} \tag{20}$$

where TP and TN indicate the count of accurately identified lesion pixels and background pixels, respectively. FP refers to the count of background pixels that are mistakenly classified as lesion pixels, whereas FN refers to the count of lesion pixels that are mistakenly classified as background pixels.

4.4. Ablation studies

For the purpose of comparing the performance of each module, we conduct ablation studies to illustrate the effectiveness of various components in the DSU-Net. During the experiments, we setup a batch size of 2 and ensured that all participants operated in identical computer environments and utilized the same data augmentation method to guarantee fairness in the comparison. First, we use RefineSegNet without the MFFM module as the baseline. Then, we gradually add MFFM modules, RoughSwinSegNet, and two-stage balanced loss to the network to construct several different experimental schemes. Finally, the ISIC2017 and ISIC2018 datasets are used to test how well every module in our method is effective.

As Table 1 illustrates, our approach increased IoU and Dice by 3.04% and 2.46%, respectively, over the baseline network after adding the MFFM module, on the ISIC 2017 dataset. These results demonstrate that MFFM can enhance the network's ability to detect and segment lesions of different scales by extracting multiple features and fusing global and local multi-scale features. Compared to RefineSegNet, which only includes the MFFM module, our method, with the addition of Swin Transformer (i.e., RoughSwinSegNet), further improved the IoU and Dice by 0.62% and 0.23%, respectively. Additionally, our method showed an increase of 0.64% in IoU and 0.47% in Dice compared to RoughSwinSegNet. This fully confirms that performing rough segmentation first and then refining the results by simulating manual segmentation is more effective than direct segmentation. With the addition of the two-stage balanced loss, our method's IoU and Dice increased by 0.63% and 0.65%, respectively, compared to not using this loss function. This can be seen as the fact that our method can better simulate the segmentation process and hence improve the segmentation performance of the model, guided by the two-stage loss function. In summary, our method outperforms the baseline network on the ISIC 2017 dataset, enhancing ACC, IoU, and Dice by 1.53%, 4.29%, and 3.34%, respectively. Similar trials on the ISIC 2018 dataset show these improvements, further verifying the effectiveness of the presented modules.

Simultaneously, to evaluate the effectiveness of each module in MFFM, we conducted ablation experiments on four modules (CAM, LAM, GAM, and MSFE) using the ISIC2017 dataset, with the results shown in Table 2. After adding the CAM module, the metrics improved significantly, demonstrating that this module effectively enhances the model's ability to perceive important features, thereby improving segmentation performance. The LAM and GAM modules are designed to extract local and global information, respectively, helping the model capture fine lesion features and strengthen its overall understanding of the lesion area, thus enhancing its ability to identify lesion regions in complex scenarios. The experiments showed that after adding the

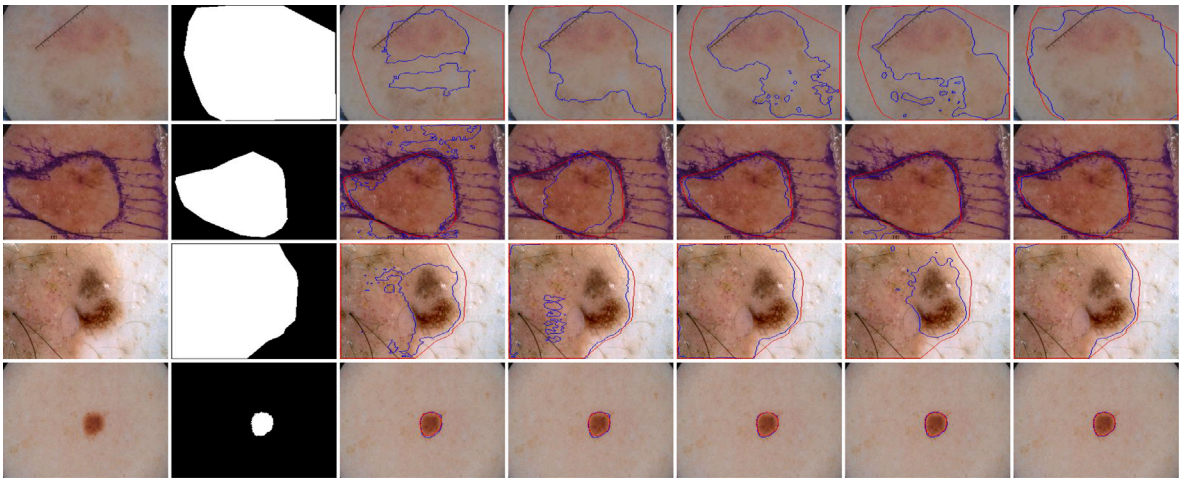

Fig. 9. A visual representation of the ablative analysis conducted on the core modules in DSU-Net. (a) Input images. (b) GroundTruth. (c)Baseline. (d)Baseline+MFFM (RefineSegNet). (e) Baseline+Swin Transformer (RoughSwinSegNet). (f) Baseline+MFFM+Swin Transformer. (g) Baseline+MFFM+Swin Transformer+Two-stage balanced loss. The red border and blue border represent the ground truth and segmentation findings, respectively.

Table 1
Ablation studies on ISIC 2017 and ISIC 2018 datasets.

Method				ISIC 2017			ISIC 2018		
Baseline	MFFM	Swin Transformer	Two-stage balanced loss	ACC (%)	IoU (%)	Dice (%)	ACC (%)	IoU (%)	Dice (%)
✓				92.51	74.20	83.16	91.44	77.18	85.45
✓	✓			93.45	77.24	85.62	93.05	80.57	87.91
✓		✓		93.96	77.22	85.38	93.92	82.41	89.33
✓	✓	✓		93.85	77.86	85.85	94.14	83.18	89.82
✓	✓	✓	✓	**94.04**	**78.49**	**86.50**	**94.31**	**83.43**	**90.04**

Table 2
Ablation study based on MFFM blocks on ISIC 2017 dataset.

Method					ISIC 2017		
Without module	CAM	LAM	GAM	MSFE	ACC (%)	IoU (%)	Dice (%)
✓					93.50	76.50	84.72
	✓				93.57	77.04	85.24
	✓	✓			93.92	77.29	85.47
	✓	✓		✓	93.95	77.50	85.61
	✓		✓		93.92	77.50	85.60
	✓		✓	✓	93.93	77.56	85.65
	✓	✓	✓	✓	**94.04**	**78.49**	**86.50**

Table 3
Comparison between RoughSwinSegNet and RefineSegNet.

Method	ISIC 2018			ISIC 2017		
	ACC (%)	IoU (%)	Dice (%)	ACC (%)	IoU (%)	Dice(%)
RoughSwinSegNet	94.26	83.19	89.91	94.00	78.28	86.37
RefineSegNet	94.31	83.43	90.04	94.04	78.49	86.50

LAM and GAM modules, the metrics increased to varying degrees. The MSFE module aims to improve the network's ability to recognize lesions of different scales, allowing the model to more comprehensively understand the shape and structure of lesions. As seen from the Table 2, adding the MSFE module also had a positive impact on the model's performance. When both local and global feature extraction branches were integrated, the performance improvement was most significant, with the Dice coefficient increasing by 0.85% and 0.89%, respectively, compared to when they were not integrated.

Furthermore, we used the ISIC2017 dataset to visualize the experimental results obtained with the aforementioned methodologies, allowing us to undertake a deeper analysis of the results. As illustrated in Fig. 9, it can be shown from the visualization results that after incorporating the MFFM module, RoughSwinSegNet, and two-stage balanced loss function, our method has significant improvements in solving complex segmentation challenges. Compared to other methods that only fuse section modules on the baseline, the DSU-Net model that combines all modules performs significantly better than these methods. At the same time, it can be found that the combined result of Baseline+MFFM+Swin Transformer is not as good as the effect of single-stage segmentation. This may be because the loss weights in the two stages are the same and the corresponding functions cannot be fully realized in the envisioned manner, resulting in insufficient rough segmentation results, which in turn affect subsequent segmentation.

Therefore, we designed a two-stage balanced loss function. By assigning a larger loss weight to the second-stage refined segmentation, the network can be guided to cover the entire target area obtained by the first-stage coarse segmentation. If the initial coarse segmentation result is insufficient, it will cause more losses in the second stage. The optimization function will guide the network to continuously expand the learned coarse segmentation area until it contains or exceeds the entire segmentation target. From the visualization results, it can be seen that after introducing the two-stage balanced loss function to control the training of the network, the performance of the combined model of Baseline+MFFM+Swin Transformer is effectively improved, thus avoiding the problem of unsatisfactory coarse segmentation results.

In addition, to visually observe the feature areas that RoughSwinSegNet and RefineSegNet may focus on, we visualize the output features of the two sub-networks at the final stage. As depicted in Fig. 10, the output features of RefineSegNet are more concentrated in the lesion area than those of RoughSwinSegNet and are more refined. We compared the results of the two-stage networks, RoughSwinSegNet and RefineSegNet, as presented in Table 3. Combining this with Table 1, we can see that through the collaborative optimization between the two stages, the segmentation results of RoughSwinSegNet have significantly improved. RefineSegNet not only plays a collaborative role but also

Table 4
Comparative statistical analysis of various state-of-the-art methodologies on the ISIC 2018 dataset.

Model	Year	SE (%)	SP (%)	ACC (%)	IoU (%)	Dice (%)	Params(M)	GFLOPs	Average time (s)
U-Net [9]	2015	87.45	93.91	90.25	73.61	82.19	7	18	0.0012
CPFNet [52]	2020	91.57	92.27	91.99	78.58	86.32	43	16	0.0040
FAT-Net [21]	2021	90.08	95.73	92.50	79.61	87.15	30	23	0.0059
MSCA-Net [37]	2023	94.34	91.85	93.05	79.86	87.86	27	12	0.0041
H2Former [22]	2023	**94.99**	89.67	92.57	80.99	88.23	33	33	0.0070
GFANet [53]	2023	93.79	91.30	92.64	79.87	87.66	24	7	0.0094
DA-TransUNet [54]	2024	93.28	94.73	93.81	81.79	89.01	108	60	0.0100
I^2UNet [55]	2024	91.12	94.84	92.96	80.63	87.97	29	9	0.0062
CSAP-UNet [56]	2024	94.51	92.40	93.42	82.01	89.13	27	9	0.0070
HTC-Net [57]	2024	91.78	94.75	92.92	79.91	87.73	116	25	0.0093
DSU-Net (ours)	–	92.22	**96.14**	**94.31**	**83.43**	**90.04**	35	19	0.0086

Table 5
Comparative statistical analysis of various state-of-the-art methodologies on the ISIC 2017 dataset.

Model	SE (%)	SP (%)	ACC (%)	IoU (%)	Dice (%)
U-Net [9]	81.55	96.36	91.45	71.83	81.23
CPFNet [52]	80.57	97.30	93.20	75.74	84.25
FAT-Net [21]	83.92	97.25	93.26	76.53	85.00
MSCA-Net [37]	82.80	96.38	93.00	76.65	84.85
H2Former [22]	82.17	96.93	93.03	76.11	84.61
GFANet [53]	82.34	96.07	93.37	76.51	84.51
DA-TransUNet [54]	81.48	**97.77**	93.35	76.31	84.83
I^2UNet [55]	83.17	96.79	93.08	76.30	84.76
CSAP-UNet [56]	81.57	97.76	93.52	76.39	84.77
HTC-Net [57]	84.69	97.27	93.52	77.13	85.22
DSU-Net (ours)	**84.70**	96.75	**94.04**	**78.49**	**86.50**

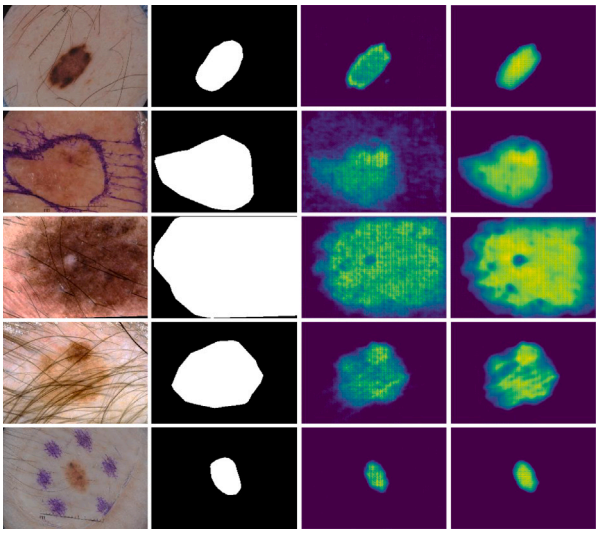

Fig. 10. Comparative analysis of the output features of the final layer in RoughSwinSegNet and RefineSegNet. (a) Input images. (b) GroundTruth. (c) RoughSwinSegNet final layer feature map. (d) RefineSegNet final layer feature map.

further enhances segmentation performance, demonstrating its ability to refine the segmentation results. This indicates that our method can better leverage the advantages of both stages to achieve precise segmentation of skin lesions.

4.5. Results on the ISIC 2018 dataset

We evaluate the DSU-Net against ten advanced approaches, namely UNet [9], CPFNet [52], FAT-Net [21], MSCA-Net [37], H2Former [22], GFANet [53], I^2UNet [55], DA-TransUNet [54], CSAP-UNet [56], and HTC-Net [57], using the ISIC 2018 dataset. Among them, FAT-Net, MSCA-Net, and GFANet are especially designed for the purpose of segmenting skin lesions, while the remaining methods are the most recent networks used for segmenting medical images. During the experiments, all participants operate within an identical computing environment and employ identical methods for data enhancement to achieve an equitable comparison. Table 4 displays the test outcomes of different methods for segmenting skin lesions using the ISIC 2018 dataset. "Average time" refers to the average time required by the model to predict a single image. FAT-Net seamlessly combines CNN with Transformer branches through a dual-branch encoder structure, achieving better performance than U-Net. H2Former integrates the local information and multi-scale channel capabilities of CNN with the long-distance features of Transformer in a unified block, further improving segmentation performance. I^2UNet adopts a dual-path structure, one path focuses on extracting image feature information, and the other path utilizes shared convolutional kernels to process hidden state information along the depth dimension. This interaction between the two paths encourages the reuse and re-exploration of historical information, allowing deep learning to integrate low-level detail descriptions with high-level semantic abstractions, thereby forming a more comprehensive feature representation and significantly enhancing segmentation performance. It is obvious that our methodology obtains the highest scores in the majority of indicators, with scores of 96.14%, 94.31%, 83.43%, and 90.04% in SP, ACC, IoU, and Dice, respectively. Among these, the model we propose ranks first on two major indices, ACC and Dice, indicating DSU-Net's competitiveness.

Moreover, we show the segmentation outcomes for different competitors through several typical cases for visual comparison. In the analyses, we have chosen four of the most representative visual comparison methods: CPFNet, FAT-Net, H2Former, and HTC-Net, as illustrated in Fig. 11. It is visible that our approach consistently outperforms its competitors and yields the most favorable segmentation outcomes in challenging instances. Among these methods, although CPFNet, FAT-Net, and HTC-Net can identify skin lesion areas, they cannot accurately predict skin lesions in challenging samples with blurred boundaries and low contrast. Although H2Former outperforms other models in segmentation outcomes, it still lacks precision in edge segmentation. In contrast, our approach attains the most optimal outcomes in the segmentation of skin lesions. Even with blurred boundaries and the existence of interference, DSU-Net can accurately segment skin lesions of varying sizes and irregular forms. This demonstrates that the two-stage segmentation method from coarse to fine can effectively remove background areas and reduce interference factors. At the same time, DSU-Net performs well in dealing with skin lesions of different shapes, which further proves that the MFFM module enhances RefineSegNet's ability to detect and understand lesion areas of different scales.

4.6. Results on the ISIC 2017 dataset

Like the ISIC 2018 dataset, we conducted a comparison between DSU-Net and ten existing algorithms using the ISIC 2017 dataset. The statistical comparison findings of these methods are presented in Table 5. CSAP-UNet is a Convolution and Self-Attention Parallel Network. It effectively merges dual-branch features and enhances

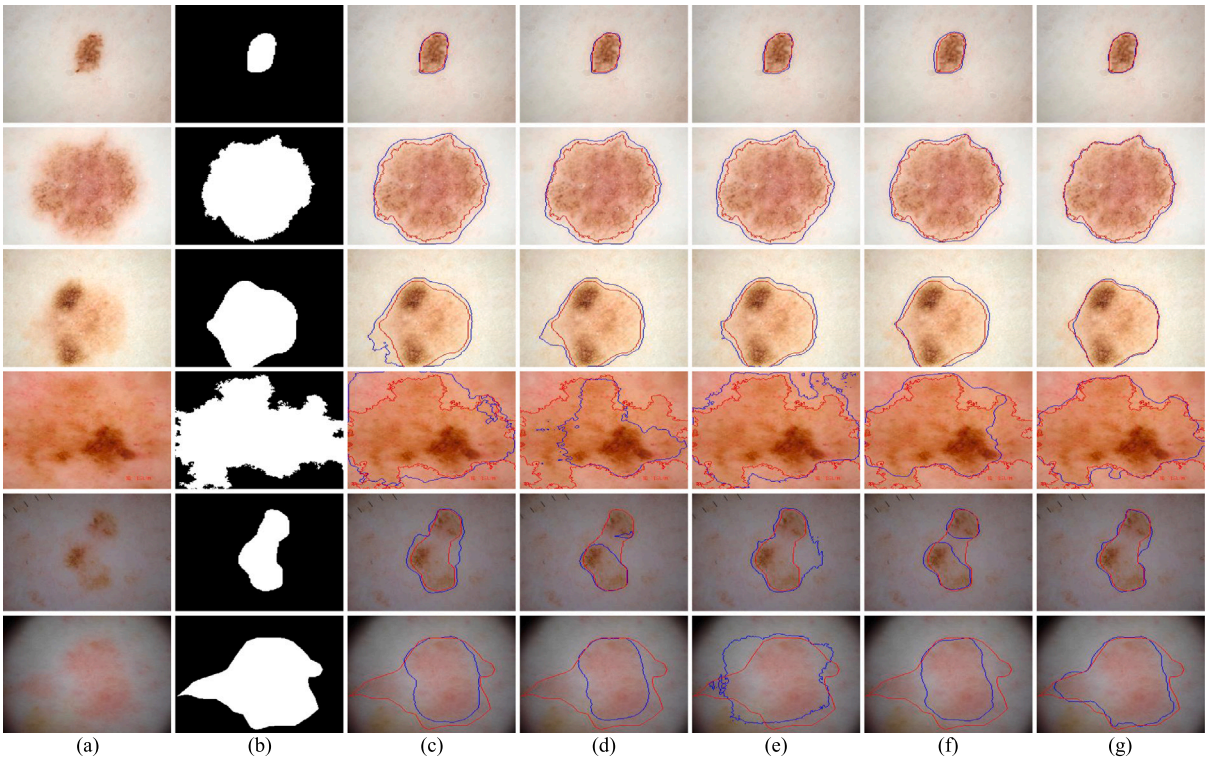

Fig. 11. Comparative analysis of various newest methods on the ISIC 2018 dataset. (a) Input images. (b) Ground truth. (c) CPF-Net [9]. (d) FAT-Net [21]. (e) H2Former [22]. (f) HTC-Net [57]. (g) DSU-Net. The red border and blue border represent the ground truth and segmentation findings, respectively.

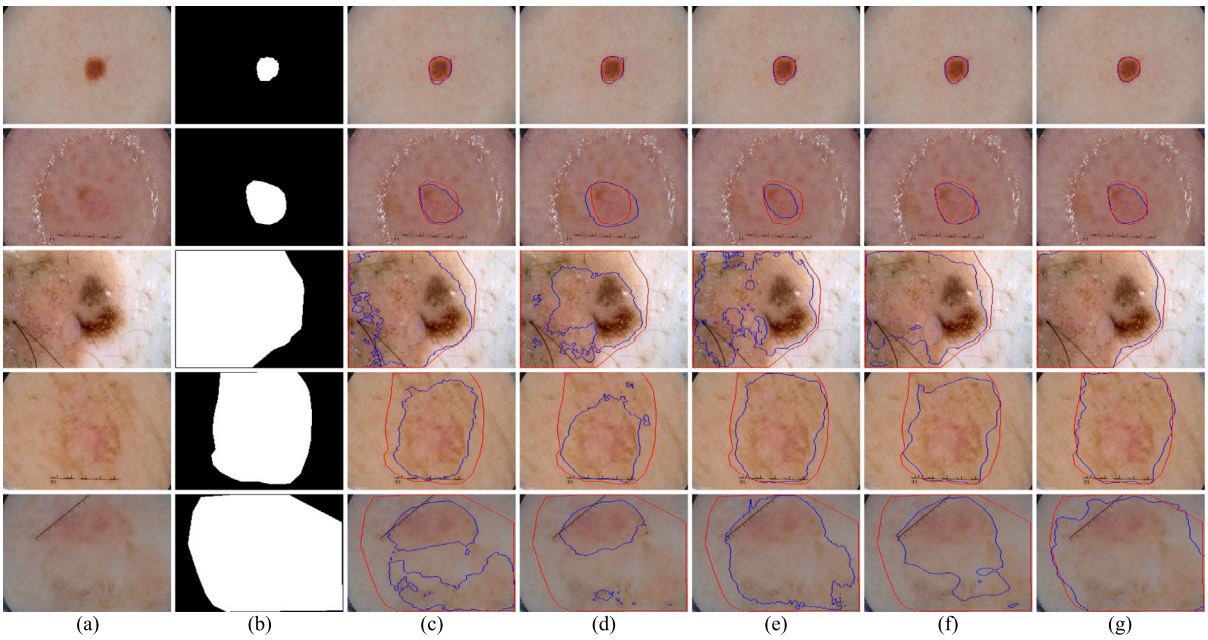

Fig. 12. Comparative analysis of various newest methods on the ISIC 2017 dataset. (a) Input images. (b) Ground truth. (c) CPF-Net [9]. (d) FAT-Net [21]. (e) H2Former [22]. (f) HTC-Net [57]. (g) DSU-Net. The red border and blue border represent the ground truth and segmentation findings, respectively.

the learning ability for target boundaries through the Attention Fusion Module (AFM) and the Boundary Enhancement Module (BEM). DSU-Net enhances ACC, IoU, and Dice by 0.52%, 2.1%, and 1.73%, respectively, in comparison to the CSAP-UNet. Even compared with the best-performing HTC-Net, DSU-Net still improves ACC, IoU, and Dice by 0.52%, 1.36%, and 1.28%, respectively. Similarly, DSU-Net typically outperforms other rivals in most accuracy and efficiency evaluation metrics on the ISIC 2017 dataset.

Further, we conduct visual comparisons of several approaches on numerous challenging examples in the ISIC 2017 dataset. As depicted in Fig. 11. Among these approaches, FAT-Net, H2Former, and HTC-Net are hybrid CNN-Transformer networks that effectively integrate local and global contextual information. However, they cannot fully predict skin lesion areas when dealing with challenging samples that have blurred boundaries and are difficult to distinguish. Compared to these competing methods, our approach can first identify the rough lesion

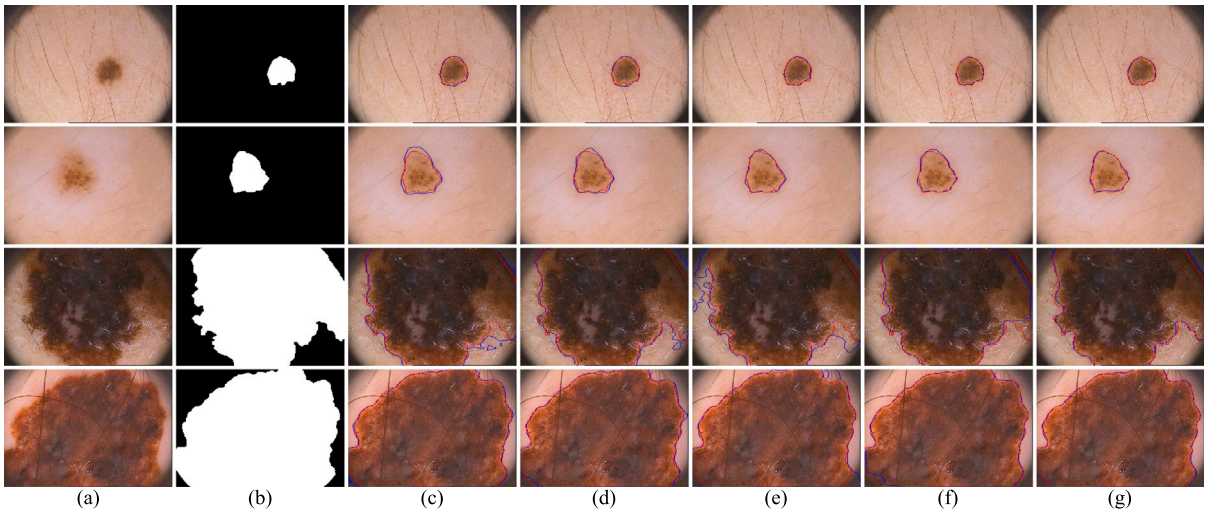

Fig. 13. Comparative analysis of various newest methods on the PH2 dataset. (a) Input images. (b) Ground truth. (c) CPF-Net [9]. (d) FAT-Net [21]. (e) H2Former [22]. (f) HTC-Net [57]. (g) DSU-Net. The red border and blue border represent the ground truth and segmentation findings, respectively.

Table 6
Comparative statistical analysis of various state-of-the-art methodologies on the PH2 dataset.

Model	SE (%)	SP (%)	ACC (%)	IoU (%)	Dice (%)
U-Net [9]	94.20	92.38	93.97	85.05	91.20
CPFNet [52]	96.75	96.46	97.49	91.69	95.56
FAT-Net [21]	**96.91**	97.25	97.57	91.73	95.60
MSCA-Net [37]	96.75	94.81	97.17	91.00	95.15
H2Former [22]	95.76	94.55	96.86	91.27	95.33
GFANet [53]	96.41	95.14	97.16	91.59	95.54
DA-TransUNet [54]	96.87	96.05	97.38	91.71	95.57
I^2UNet [55]	94.71	95.93	96.61	89.24	94.21
CSAP-UNet [56]	96.13	97.29	97.66	92.16	95.84
HTC-Net [57]	95.89	96.50	97.32	91.27	95.37
DSU-Net (ours)	96.27	**97.98**	**97.80**	**92.55**	**96.04**

region to remove some background interference and then perform precise boundary segmentation, thereby more effectively segmenting skin lesions. As evident from Fig. 12, the segmentation results of DSU-Net are superior and closely resemble the actual scenario.

4.7. Results on the PH2 dataset

Ultimately, we perform tests of contrast on the PH2 dataset to showcase the stability of the presented DSU-Net. We adopt the identical comparison approach utilized in ISIC 2017 and ISIC 2018, where we compare our method with ten other methods: UNet [9], CPFNet [52], FAT-Net [21], MSCA-Net [37], H2Former [22], GFANet [53], DA-TransUNet [54], I^2UNet [55], CSAP-UNet [56], and HTC-Net [57]. As shown in Table 6, our DSU-Net surpasses other rivals overall, achieving 97.80% in ACC, 92.55% in IoU, and 96.04% in Dice.

Simultaneously, we also do a visual analysis of various approaches on multiple typical challenging instances in the PH2 dataset. Four exemplary visual comparison approaches are chosen for comparison in the experiments. As depicted in Fig. 13, our method consistently achieves superior segmentation results in difficult scenarios characterized by intricate borders and changing brightness. This clearly demonstrates the proposed DSU-Net generalization capability and effectiveness.

5. Discussion and limitations

To better segment skin diseases, we present a dual-stage U-Net (DSU-Net) based on CNN and Transformers. This method draws on the manual segmentation process and divides the task into two steps. First, the coarse lesion area is found using RoughSwinSegNet, and then the boundaries of these areas are refined by RefineSegNet to achieve accurate segmentation. We use Swin Transformer as the encoder of RoughSwinSegNet for feature extraction and load its pre-trained model to better identify the approximate lesion area. To enhance the boundary recognition ability of RefineSegNet, we introduce the MFFM module, which enhances DSU-Net's ability to detect and understand lesion areas at different scales by fusing global and local multi-scale features. To avoid the coarse segmentation outcome's inadequacy, we designed a dual-stage segmentation loss function to control the loss weights of the two networks and strengthen their synergy. As a result, RefineSegNet's segmentation output is more accurate, compensating for RoughSwinSegNet's limitations. This method can eliminate interference factors such as blood vessels and hair before performing accurate recognition. The experimental results demonstrate that our method expresses superior performance in comparison to direct segmentation.

While DSU-Net demonstrates strong performance on three publicly accessible skin lesion datasets, like many other powerful segmentation algorithms, our network still has limitations when it comes to handling low contrast. As shown in Fig. 14, when the contrast between the lesion area and the surrounding tissue in the skin image is very low, or when there are significant color variations within the lesion, our method fails to accurately segment the lesion boundaries. Additionally, for these complex and challenging samples, our RoughSwinSegNet also has difficulty accurately finding the coarse lesion area, resulting in insufficient coarse segmentation results. Given the recent emergence of Mamba [58] technology, we plan to investigate a hybrid network architecture integrating Mamba and Transformer for medical image segmentation in future studies. More precisely, we will utilize the benefits of both Mamba and Transformer to create a novel two-stage segmentation network structure to tackle the issue of low-contrast problematic samples.

6. Conclusion

Inspired by the manual segmentation process, we propose a dual-stage U-Net (DSU-Net) based on CNN and Transformers for automatic skin lesion, which can achieve accurate segmentation by simulating a coarse-to-fine segmentation process. In the first stage, the rough lesion regions are identified by RoughSwinSegNet. In the second stage, the boundaries of these regions are refined by RefineSegNet. In addition, a two-stage loss function is specifically designed to guide the network to

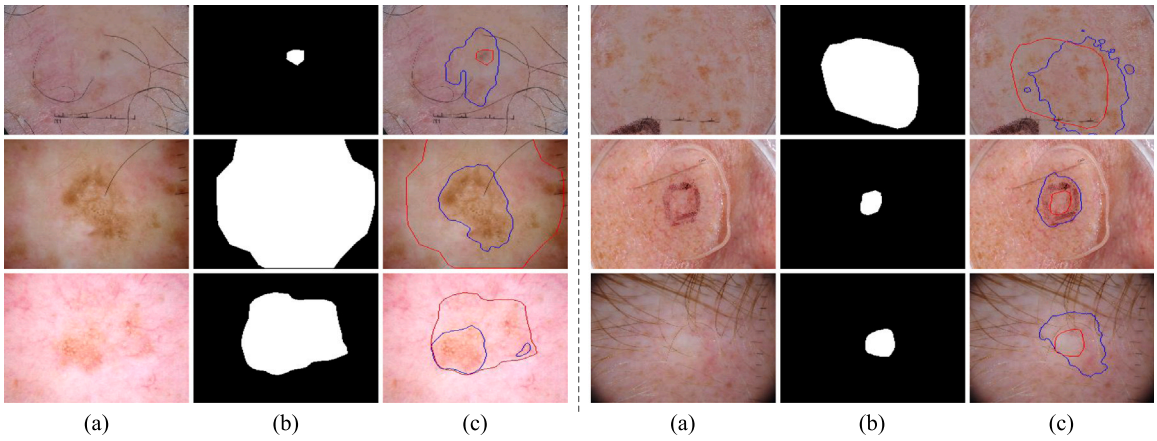

Fig. 14. Visualization of DSU-Net's poor segmentation performance on the ISIC2017 and ISIC2018 datasets. (a) Input images. (b) Ground truth. (c) DSU-Net. The red border and blue border represent the ground truth and segmentation findings, respectively.

adaptively learn this process, thereby improving the segmentation accuracy. Extensive experiments demonstrate that our DSU-Net outperforms comparison methods on three public datasets, especially when dealing with complex boundaries and low-contrast samples. Intuitive visualization results further show that our DSU-Net performs more satisfactorily in addressing various skin disease segmentation challenges.

CRediT authorship contribution statement

Longwei Zhong: Writing – review & editing, Software, Methodology, Conceptualization. **Tiansong Li:** Writing – review & editing, Supervision, Funding acquisition. **Meng Cui:** Writing – review & editing, Visualization, Formal analysis. **Shaoguo Cui:** Writing – review & editing, Supervision. **Hongkui Wang:** Writing – review & editing, Supervision. **Li Yu:** Writing – review & editing, Supervision.

Declaration of competing interest

We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contributions to this work. We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property. We further confirm that any aspect of the work covered in this manuscript that has involved either experimental animals or human patients has been conducted with the ethical approval of all relevant bodies and that such approvals are acknowledged within the manuscript.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62202134, in part by the Natural Science Foundation Project of Chongqing Science and Technology Bureau (CSTB2022NSCQ- MSX1231), in part by the Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN20220051), in part by the Talents Fund Project of Chongqing Normal University (21XLB031), in part by the Pioneer and Leading Goose R and D Program of Zhejiang Province under Grant 2023C01149, and in part by the Graduate Research Innovation Project of Chongqing Normal University(YKC23034).

Data availability

The data and code used in this study are openly available at the following repository: Name: DSU-Net URL: https://github.com/ZhongLongwei/DSU-Net.

References

[1] R.L. Siegel, K.D. Miller, N.S. Wagle, A. Jemal, Cancer statistics, 2023, Ca Cancer J. Clin. 73 (1) (2023) 17–48.
[2] L.K. Huang, M.J.J. Wang, Image thresholding by minimizing the measures of fuzziness, Pattern Recognit. 28 (1) (1995) 41–51.
[3] P. Schmid, Segmentation of digitized dermatoscopic images by two-dimensional color clustering, IEEE Trans. Med. Imaging 18 (2) (1999) 164–171.
[4] M.E. Celebi, S. Hwang, H. Iyatomi, G. Schaefer, Robust border detection in dermoscopy images using threshold fusion, in: 2010 IEEE International Conference on Image Processing, IEEE, 2010, pp. 2541–2544.
[5] G. Sforza, G. Castellano, S.K. Arika, R.W. LeAnder, R.J. Stanley, W.V. Stoecker, J.R. Hagerty, Using adaptive thresholding and skewness correction to detect gray areas in melanoma in situ images, IEEE Trans. Instrum. Meas. 61 (7) (2012) 1839–1847.
[6] H. Zhou, M. Chen, L. Zou, R. Gass, L. Ferris, L. Drogowski, J.M. Rehg, Spatially constrained segmentation of dermoscopy images, in: 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano To Macro, IEEE, 2008, pp. 800–803.
[7] H. Zhou, G. Schaefer, A.H. Sadka, M.E. Celebi, Anisotropic mean shift based fuzzy c-means segmentation of dermoscopy images, IEEE J. Sel. Top. Sign. Proces. 3 (1) (2009) 26–34.
[8] A. Wong, J. Scharcanski, P. Fieguth, Automatic skin lesion segmentation via iterative stochastic region merging, IEEE Trans. Inf. Technol. Biomed. 15 (6) (2011) 929–936.
[9] O. Ronneberger, P. Fischer, T. 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, Springer, 2015, pp. 234–241.
[10] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2019) 1856–1867.
[11] Z. Zhang, Q. Liu, Y. Wang, Road extraction by deep residual u-net, IEEE Geosci. Remote Sens. Lett. 15 (5) (2018) 749–753.
[12] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[13] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, 2018, arXiv preprint arXiv:1804.03999.

[14] L. Yu, H. Chen, Q. Dou, J. Qin, P.-A. Heng, Automated melanoma recognition in dermoscopy images via very deep residual networks, IEEE Trans. Med. Imaging 36 (4) (2016) 994–1004.
[15] P. Tang, Q. Liang, X. Yan, S. Xiang, W. Sun, D. Zhang, G. Coppola, Efficient skin lesion segmentation using separable-unet with stochastic weight averaging, Comput. Methods Programs Biomed. 178 (2019) 289–301.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[18] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.
[19] H.Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, Nnformer: Interleaved transformer for volumetric segmentation, 2021, arXiv preprint arXiv:2109.03201.
[20] X. Huang, Z. Deng, D. Li, X. Yuan, Y. Fu, Missformer: An effective transformer for 2d medical image segmentation, IEEE Trans. Med. Imaging (2022).
[21] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, Z. Wen, FAT-net: Feature adaptive transformers for automated skin lesion segmentation, Med. Image Anal. 76 (2022) 102327.
[22] A. He, K. Wang, T. Li, C. Du, S. Xia, H. Fu, H2former: An efficient hierarchical hybrid transformer for medical image segmentation, IEEE Trans. Med. Imaging (2023).
[23] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[24] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 801–818.
[25] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[26] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu, Ce-net: Context encoder network for 2d medical image segmentation, IEEE Trans. Med. Imaging 38 (10) (2019) 2281–2292.
[27] L. Yang, C. Fan, H. Lin, Y. Qiu, MaMfi-Net: Multi-attention and multi-feature interaction network in skin lesion segmentation, Biomed. Signal Process. Control 96 (2024) 106567.
[28] W. Zhu, J. Tian, M. Chen, L. Chen, J. Chen, MSS-UNet: A multi-spatial-shift MLP-based unet for skin lesion segmentation, Comput. Biol. Med. 168 (2024) 107719.
[29] Y. Yuan, M. Chao, Y.-C. Lo, Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance, IEEE Trans. Med. Imaging 36 (9) (2017) 1876–1886.
[30] Y. Xie, J. Zhang, Y. Xia, C. Shen, A mutual bootstrapping model for automated skin lesion segmentation and classification, IEEE Trans. Med. Imaging 39 (7) (2020) 2482–2493.
[31] F. Bagheri, M.J. Tarokh, M. Ziaratban, Skin lesion segmentation from dermoscopic images by using mask R-CNN, retina-deeplab, and graph-based methods, Biomed. Signal Process. Control 67 (2021) 102533.
[32] D. Dai, C. Dong, S. Xu, Q. Yan, Z. Li, C. Zhang, N. Luo, Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation, Med. Image Anal. 75 (2022) 102293.
[33] K. Hu, J. Lu, D. Lee, D. Xiong, Z. Chen, AS-net: Attention synergy network for skin lesion segmentation, Expert Syst. Appl. 201 (2022) 117112.
[34] H. Wu, J. Pan, Z. Li, Z. Wen, J. Qin, Automated skin lesion segmentation via an adaptive dual attention module, IEEE Trans. Med. Imaging 40 (1) (2020) 357–370.
[35] J. Ruan, S. Xiang, M. Xie, T. Liu, Y. Fu, Malunet: A multi-attention and light-weight unet for skin lesion segmentation, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2022, pp. 1150–1156.
[36] R. Arora, B. Raman, K. Nayyar, R. Awasthi, Automated skin lesion segmentation using attention-based deep convolutional neural network, Biomed. Signal Process. Control 65 (2021) 102358.
[37] Y. Sun, D. Dai, Q. Zhang, Y. Wang, S. Xu, C. Lian, MSCA-net: Multi-scale contextual attention network for skin lesion segmentation, Pattern Recognit. 139 (2023) 109524.
[38] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.
[39] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.
[40] X. He, E.L. Tan, H. Bi, X. Zhang, S. Zhao, B. Lei, Fully transformer network for skin lesion analysis, Med. Image Anal. 77 (2022) 102357.
[41] X. Liu, P. Gao, T. Yu, F. Wang, R.Y. Yuan, CSWin-UNet: Transformer UNet with cross-shaped windows for medical image segmentation, Inf. Fusion (2024) 102634.
[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[43] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, J. Qin, Boundary-aware transformers for skin lesion segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 206–216.
[44] W. Cao, G. Yuan, Q. Liu, C. Peng, J. Xie, X. Yang, X. Ni, J. Zheng, ICL-Net: Global and local inter-pixel correlations learning network for skin lesion segmentation, IEEE J. Biomed. Health Inf. 27 (1) (2022) 145–156.
[45] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, D. Zhang, Ds-transunet: Dual swin transformer u-net for medical image segmentation, IEEE Trans. Instrum. Meas. 71 (2022) 1–15.
[46] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, et al., TransUNet: Rethinking the U-net architecture design for medical image segmentation through the lens of transformers, Med. Image Anal. 97 (2024) 103280.
[47] Z. Huang, H. Deng, S. Yin, T. Zhang, W. Tang, Q. Wang, ADF-Net: A novel adaptive dual-stream encoding and focal attention decoding network for skin lesion segmentation, Biomed. Signal Process. Control 91 (2024) 105895.
[48] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
[49] N.C. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al., Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 168–172.
[50] N. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019, arXiv preprint arXiv:1902.03368.
[51] T. Mendonça, P.M. Ferreira, J.S. Marques, A.R. Marcal, J. Rozeira, PH 2-a dermoscopic image database for research and benchmarking, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, IEEE, 2013, pp. 5437–5440.
[52] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, X. Chen, CPFNet: Context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imaging 39 (10) (2020) 3008–3018.
[53] S. Qiu, C. Li, Y. Feng, S. Zuo, H. Liang, A. Xu, GFANet: Gated fusion attention network for skin lesion segmentation, Comput. Biol. Med. 155 (2023) 106462.
[54] G. Sun, Y. Pan, W. Kong, Z. Xu, J. Ma, T. Racharak, L.M. Nguyen, J. Xin, DA-TransUNet: integrating spatial and channel dual attention with transformer U-net for medical image segmentation, Front. Bioeng. Biotechnol. 12 (2024) 1398237.
[55] D. Dai, C. Dong, Q. Yan, Y. Sun, C. Zhang, Z. Li, S. Xu, I2U-Net: A dual-path U-Net with rich information interaction for medical image segmentation, Med. Image Anal. (2024) 103241.
[56] X. Fan, J. Zhou, X. Jiang, M. Xin, L. Hou, CSAP-UNet: Convolution and self-attention paralleling network for medical image segmentation with edge enhancement, Comput. Biol. Med. 172 (2024) 108265.
[57] H. Tang, Y. Chen, T. Wang, Y. Zhou, L. Zhao, Q. Gao, M. Du, T. Tan, X. Zhang, T. Tong, HTC-Net: A hybrid CNN-transformer framework for medical image segmentation, Biomed. Signal Process. Control 88 (2024) 105605.
[58] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, 2023, arXiv preprint arXiv:2312.00752.