Computer Methods and Programs in Biomedicine 238 (2023) 107601

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Dynamically aggregating MLPs and CNNs for skin lesion segmentation with geometry regularization

Chuanbo Qin[a], Bin Zheng[a], Junying Zeng[a,*], Zhuyuan Chen[a], Yikui Zhai[a], Angelo Genovese[b], Vincenzo Piuri[b], Fabio Scotti[b]

[a] *Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen 529020, China*
[b] *Departimento di Information, Università degli Studi di Milano, 20133 Milano, Italy*

ARTICLE INFO

Article history:
Received 14 October 2022
Revised 24 April 2023
Accepted 13 May 2023

Keywords:
Skin lesion segmentation
Multi-layer perceptions
Feature interaction
Geometric information

ABSTRACT

Background and objective: Melanoma is a highly malignant skin tumor. Accurate segmentation of skin lesions from dermoscopy images is pivotal for computer-aided diagnosis of melanoma. However, blurred lesion boundaries, variable lesion shapes, and other interference factors pose a challenge in this regard.

Methods: This work proposes a novel framework called CFF-Net (Cross Feature Fusion Network) for supervised skin lesion segmentation. The encoder of the network includes dual branches, where the CNNs branch aims to extract rich local features while MLPs branch is used to establish both the global-spatial-dependencies and global-channel-dependencies for precise delineation of skin lesions. Besides, a feature-interaction module between two branches is designed for strengthening the feature representation by allowing dynamic exchange of spatial and channel information, so as to retain more spatial details and inhibit irrelevant noise. Moreover, an auxiliary prediction task is introduced to learn the global geometric information, highlighting the boundary of the skin lesion.

Results: Comprehensive experiments using four publicly available skin lesion datasets (i.e., ISIC 2018, ISIC 2017, ISIC 2016, and PH2) indicated that CFF-Net outperformed the state-of-the-art models. In particular, CFF-Net greatly increased the average Jaccard Index score from 79.71% to 81.86% in ISIC 2018, from 78.03% to 80.21% in ISIC 2017, from 82.58% to 85.38% in ISIC 2016, and from 84.18% to 89.71% in PH2 compared with U-Net. Ablation studies demonstrated the effectiveness of each proposed component. Cross-validation experiments in ISIC 2018 and PH2 datasets verified the generalizability of CFF-Net under different skin lesion data distributions. Finally, comparison experiments using three public datasets demonstrated the superior performance of our model.

Conclusion: The proposed CFF-Net performed well in four public skin lesion datasets, especially for challenging cases with blurred edges of skin lesions and low contrast between skin lesions and background. CFF-Net can be employed for other segmentation tasks with better prediction and more accurate delineation of boundaries.

1. Introduction

The incidence rates of skin cancer have shown a rapid increase across the world [1]. Melanoma is a highly malignant skin tumor associated with a high mortality rate [2]. Accurate skin lesion segmentation by dermoscopy is an effective means to conduct computer-assisted diagnosis and treatment for melanoma. Compared with the conventional manual skin lesion delineation method, automatic segmentation of skin lesion areas saves time and effort.

With the recent development of deep learning in the field of computer vision (CV), convolutional neural networks (CNNs) are being widely employed for medical image segmentation tasks. The majority of contemporary models, such as the fully convolutional network [3], U-Net [4], and U-Net++ [5], use an encoder-decoder design, wherein the encoder is applied to extract the features and the decoder is employed to restore the features to the original image size. However, accurate and efficient segmentation of skin lesions remains a challenge for these models. First, the size and shape of the skin lesions are distinct. Second, the boundaries of skin lesions are often blurred, and the lesions possess low contrast from their surrounding tissues. Third, several interference factors, such as hair, rule marks, and color calibration charts, can also affect the segmentation accuracy. Fig. 1 shows some challenging

[*] Corresponding author.
E-mail address: zengjunying@126.com (J. Zeng).

https://doi.org/10.1016/j.cmpb.2023.107601

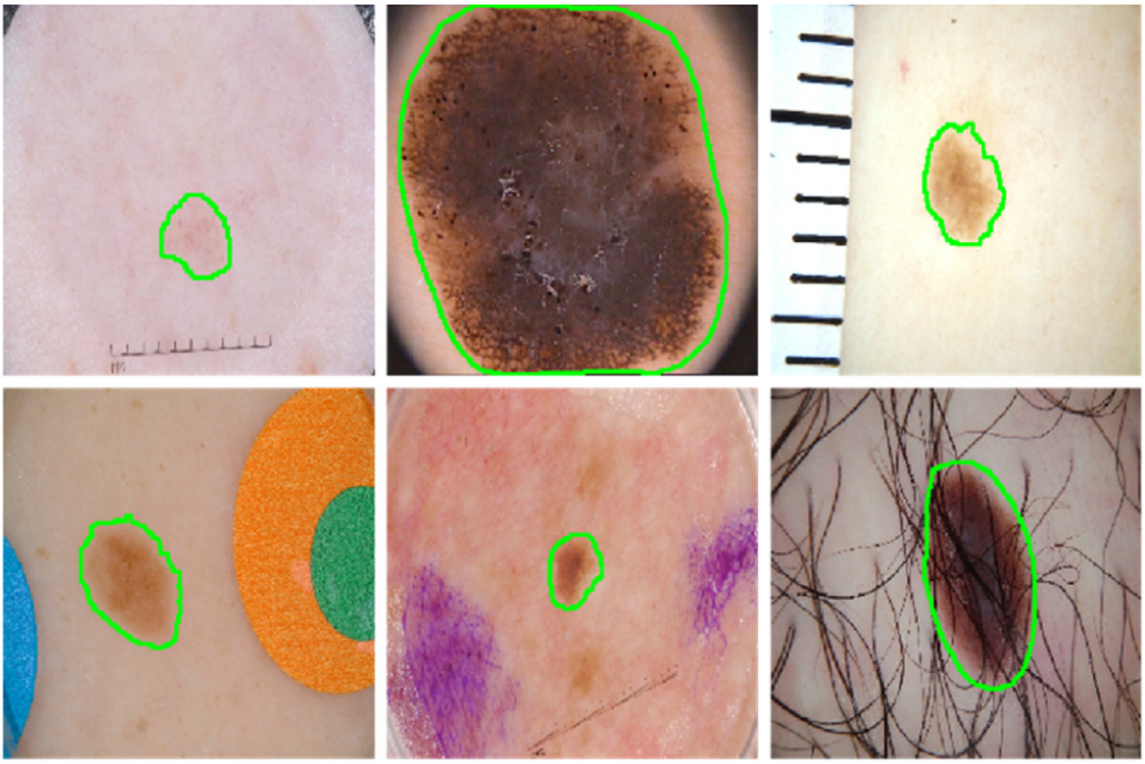

Fig. 1. Some representative examples illustrating the challenges (e.g., blurred lesion boundary with low contrast, irregular shape, and some interference factors) in ISIC 2018. The green lines in the images indicate ground truths.

cases. In recent years, a series of methods have been proposed to improve the segmentation performance for skin lesions. For example, the CA-Net [6] was fused with the spatial attention, channel attention, and scale attention to focus on the areas of interest. In addition, CPF-Net [7] was equipped with two pyramidal modules to capture global and multi-scale information. Moreover, Dai et al. (2021) [8] designed the Ms RED, which can extract rich hierarchical features and retain helpful information during down-sampling. However, the segmentation results obtained with these models are still unsatisfactory owing to their inability to capture long-range dependencies that enable precise localization of skin lesions.

More recently, the transformer, which was first proposed in the domain of natural language processing, employed the self-attention mechanism to establish long-range dependencies [9]. Unlike the CNNs used for the CV, vision transformer [10] can utilize multi-head self-attention mechanism to extract global context information, resulting in favorable image recognition performance. Researchers have also integrated the transformers into the medical image segmentation models, including TransUNet [11], Swin-Unet [12], and MedT [13]. In particular, FAT-Net [14] integrates the CNNs and transformers to complete skin lesion segmentation task. While FAT-Net can establish both local and global dependencies, it cannot dynamically integrate the global and local features, which is helpful in recovering more spatial details. In addition, the models incorporated with transformers failed to build global-channel dependencies and global-spatial dependencies synchronously [15], which limits the extraction capability of the discriminative feature. Furthermore, transformer-based architecture is constrained by a large number of parameters and a heavy dependence on training data [16], which affects the clinical practicability. The UNeXt [17], which is the first model to employ CNNs and multi-layer perceptions (MLPs) for medical image segmentation, showed great performance in segmentation and reduced the computational complexity. The combination of CNNs and MLPs in UNeXt was based on the sequential mode, which constrains the interaction of global and local features; therefore, it failed to capture enough spatial information and suppress irrelevant information for complex and fine skin lesions. To address the above issues of segmentation models, we developed a novel model for more accurate and reliable skin lesion segmentation with comparable parameters. In the network, the encoder contains dual branches (i.e., CNNs branch and MLPs branch) to capture local and global context information. In addition, we propose a feature-interaction module based on cross-attention mechanism which can enable dynamic exchange between spatial and channel information of the two branches in order to recover more spatial details and emphasize the discriminative features. In the decoding phase, the skip-connection from the improved CNNs branch is utilized to compensate the information loss caused by consecutive down-sampling in the encoding phase.

Previous models applied boundary attention [18] or explored the edge loss function [19] to further highlight the boundary details of skin lesions. While providing segmentation results, these models lack global shape awareness which helps reduce the false-positives and suppress irrelevant noise. In this work, we introduce the multi-task learning strategy (i.e., dual-task heads) to guide the network for more accurate delineation of the boundaries of skin lesions. The output of our model consists of two parts, i.e., the binary segmentation map (BSM) and the signed distance map (SDM) [20,21]. The closest distance between a pixel and boundary of the skin lesion, given a pixel in image space, determines the absolute value of the SDM. In particular, the shape change can only affect the local pixels in the binary segmentation map, while this change can globally alter the values of multiple pixels of the SDM. Accurate prediction of SDM enhances the geometric awareness of the network, thus reducing errors in over-segmentation and under-segmentation of skin lesions.

In summary, we term our model as Cross Feature Fusion Network (CFF-Net). The performance of our proposed model is assessed using four public skin lesion datasets, including ISIC 2018 [22,23], ISIC 2017 [24], ISIC 2016 [25], and PH2 [26]. In addition, we demonstrate the effectiveness of each component through ablation experiments. To verify the generalizability of CFF-Net over different distributions of skin lesion data, we perform cross-validation experiments in ISIC 2018 and PH2. Furthermore, we conduct comparison experiments on other three public datasets to demonstrate the universality and robustness of CFF-Net. In short, our work mainly contributes in three aspects:

(1) A novel encoder combines CNNs branch and MLPs branch in parallel, which is conducive to the extraction of rich local features while simultaneously building global-spatial-dependencies and global-channel-dependencies. Furthermore, a new feature-interaction module is applied to refine the features of MLPs branch and CNNs branch by allowing dynamic exchange of spatial and channel information.
(2) The introduction of auxiliary loss on SDM predictions is conducted to learn the position and shape information of the skin lesions, further highlighting the boundaries of segmentation predictions.
(3) Comprehensive experiments show that CFF-Net achieves state-of-the-art segmentation performance in four public skin lesion datasets (namely ISIC 2016, ISIC 2017, ISIC 2018, PH2), while having comparable parameters. Additionally, we adopt comparison experiments on three public datasets to demonstrate the applicability of CFF-Net for other segmentation tasks.

2. Related works

2.1. MLPs module

Recently, MLP-based networks [15,26–28] have been proposed for the processing of CV tasks. As a pure MLP-based network, MLP-Mixer [27] applies token-mixing MLP and channel-mixing MLP operation to set up communication between different channels and spatial locations. The experiments demonstrated that MLP-Mixer has the same performance as the existing CNN-based and transformer-based networks, but it requires less computations. Res-MLP [15], trained only on the ImageNet-1 K, showed good classification performance with residual MLP. AS-MLP [28] introduced local information through the axially shifting feature map and S^2-MLP [29] utilized the spatial-shift operation to allow the interaction between different spatial locations, while CycleMLP [30] based on Cycle FC operator showed good performance in various dense prediction tasks. To summarize, these MLP-based networks focus on capturing channel-dependencies and spatial-dependencies synchronously while maintaining low computation resources. The first convolutional MLP model called UNeXt [17] was proposed for medical image segmentation, which showed leading-edge performance under less parameters. However, it only connected CNN-based module and MLP-based module in sequential mode, thus ignoring the interaction between the features of different modules.

2.2. Signed distance map for medical image segmentation

In recent years, several studies have explored the application of distance map in medical image segmentation, which usually represents the geometric information. Xue et al. (2020) [21] integrated SDM learning mechanism into the 3DUNet [29] for 3D organ segmentation. Wang et al. (2020) [31] proposed Deep Distance Transform (DDT) for segmentation of tubular structures in medical images; it improved the accuracy of segmentation of tubular structure targets from complex background. Moreover, some researchers combined the SDM with semi-supervised learning method; for example, Li et al. (2020) [32] developed additional SDM prediction task to render the shape representation of unlabeled data consistent with that of labeled data. Similarly, Liu et al. (2022) [33] used the geometry-aware consistency regularization based on SDM for semi-supervised segmentation. Particularly, Phan et al. [34] designed an architecture based on U-Net introducing the auxiliary task about SDM regression, to improve the localization of skin lesions. Similarly, we adopt multi-task loss for predicting the SDM and BSM, in order to better delineate the boundaries of the skin lesion. Unlike this model, we use a single convolution layer instead of an independent decoder to output the predictions for each task, in order to reduce the parameters.

2.3. Skin lesion segmentation

In recent years, deep learning methods have been widely used for analysis of skin lesions including classification [35], segmentation [36], detection [37,38], and recognition [39], because these can adaptively learn powerful features without manual intervention. Several models proposed for skin lesion segmentation have shown promising performance over the last few years. Kadry et al. [40] employed the pre-trained VGG-SegNet to extract the fragment skin melanoma from dermoscopy images. Wang et al. [41] proposed a dual objective network for skin lesion segmentation, in which the recurrent context encoding module is designed to construct the relation among multi-scale features. Wu et al. [42] presented the dual attention module for automated skin lesion segmentation, which can capture multi-scale complementary global features. Wang et al. [43] devised a cascaded context enhancement network for skin lesion segmentation, which enriches the global context and adaptively integrates local contextual information. Although these methods have improved the segmentation accuracy for skin lesions, they cannot build the global-spatial dependencies and global-channel dependencies simultaneously, which can help retain more spatial details and emphasize the discriminative features. In addition, they neglect the interaction of global and local features, which can refine semantic representation of the network. Moreover, these models do not take the geometric constraint into account, which has an impact on the localization of skin lesions with indistinct boundaries.

3. Methodology

A schematic illustration of our CFF-Net based on encoder-decoder architecture is presented in Fig. 2. The encoder consists of two parallel branches that serve to model different context information: (1) CNNs branch focuses on constructing the local-spatial-dependencies, namely extraction of local features; (2) MLPs branch is mainly responsible for capturing the global-channel-dependencies and global-spatial-dependencies. Subsequently, features of the same scale extracted from two branches pass through a novel feature-interaction module (FIM), which enables dynamic exchange of contextual information between various features to enhance the representation ability. Concatenated high-level features of CNNs and MLPs are then fed into the decoder where progressive deconvolution operations are employed to restore the size of features and enhanced skip-connections are utilized to compensate for the loss of local and global context information. Finally, the prediction-head of the network is formed using BSMs and signed distance maps SDMs.

3.1. CNNs and MLPs branch

Before passing through the dual branch encoder, the images are fed into a single CNN-based module, including two ConvBnRe layers. The CNNs branch has three CNN-based modules, and each

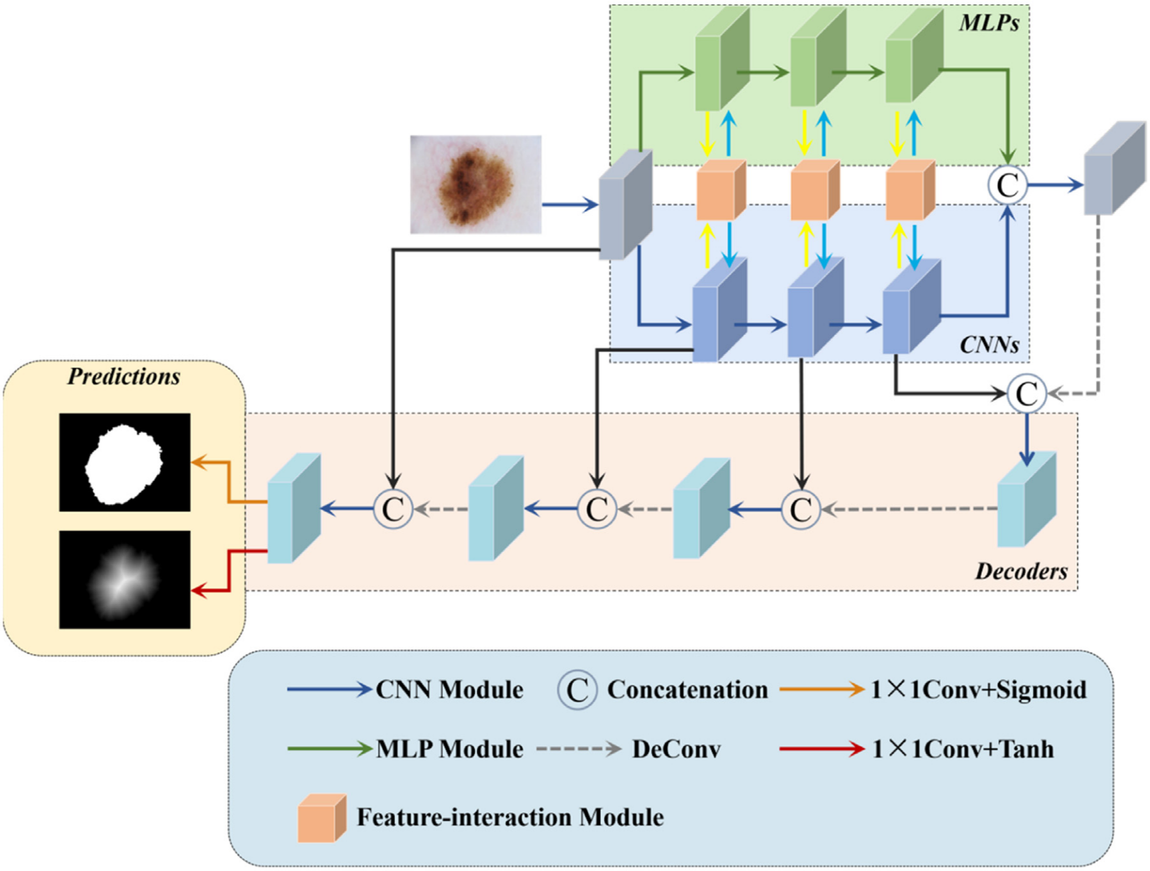

Fig. 2. Overview of the proposed CFF-Net. CNN-based module contains two ConvBnRe layers, each of which is formed by convolution layer with a kernel size of 3 × 3, batch normalization layer and Rectified Linear Units (ReLU) activation; MLP-based module contains two MLP-based operations in different directions and 1 × 1 convolution layer, as specified in Section 3.1.

module applies a max-pooling layer to down-sample the feature maps.

The MLPs branch is composed of three MLP-based modules, each of which comprises Axial-Shifted-H MLP, Axial-Shifted-W MLP, and a 1 × 1 convolution layer, as shown in Fig. 3A (left). In order to model spatial relationships in different directions, axial shift strategy [44,45] is employed to construct the MLP module. The difference between Axial-Shifted-H MLP and Axial-Shifted-W MLP lies in the direction of shift operation: one for the height direction and the other for the width direction, as displayed in Fig. 3B. For convenience, we only expound the Axial-Shifted MLP of one direction, as displayed in Fig. 3A (right).

3.1.1. Axial-shifted MLP

Given the feature map $M \in R^{C \times H \times W}$, it firstly passes through the linear embedding layer and Layer Norm (LN) [46] in Axial-Shifted MLP, which changes the original feature map into token with the size of $R^{C_e \times \frac{H}{2} \times \frac{W}{2}}$. In the four MLP-based modules, we set the size of C_e at 32, 64, 128, and 256, respectively. Then, Axial-Shifted MLP performs the shift operation on the token (see Fig. 3B), and we slice the token into $2n + 1$ groups along the channel direction, in which the shift stride for the i-th ground of token is $i - n - 1$. Subsequently, the token is reshaped and transposed to the size of $R^{\frac{HW}{4} \times C_e}$ and then goes through the Shifted-MLP block, which consists of Fully-Connected layer, depth convolution (DW-CONV) layer, GELU [47] activation, and LN. In addition, we use the Skip-Connection from the ResNet [48] to prevent gradients from vanishing. The Axial-Shifted MLP can be formulated as follows:

$$E = LN(E_{em}(M)) \tag{1}$$

$$E_{shift} = Shift(E) \tag{2}$$

$$E_{out} = f\left(RT\left(E_{shift}\right)\right) \tag{3}$$

$$Y_{out} = LN(\sigma(E_{out})) + E \tag{4}$$

where H and W are the height and width of the feature map. Note that, n is an integer greater than 1, and we set it to 2 in this paper. E_{em} denotes the 3 × 3Conv with the stride of 2, RT denotes the operation of reshaping and transposing, and *Shift* is the Shift Operation in height or width direction. Moreover, f(.) represents the cascade Fully-Connected and DW-CONV layer, and $\sigma(.)$ represents the GELU activation.

3.1.2. Fully-connected and DW-CONV layer

Fully-Connected layer is conducted for the shifted token with the size of $R^{\frac{HW}{4} \times C_e}$, which allows global communication between diverse channels in different spatial locations, as shown in Fig. 3C. Next, the token is reshaped to the size of $R^{C_e \times \frac{H}{2} \times \frac{W}{2}}$ and passed through a DW-CONV layer, in which the tokens from different groups have exclusive dilation rate due to the Shift Operation. Based on this, both local and global spatial information is encoded to enhance the feature representation, while requiring fewer parameters. Compared with the CNNs and transformers, MLPs can not only simultaneously focus on the extraction of local features and global features, but can also exploit the Shift Operation to establish global channel interaction in different spatial locations.

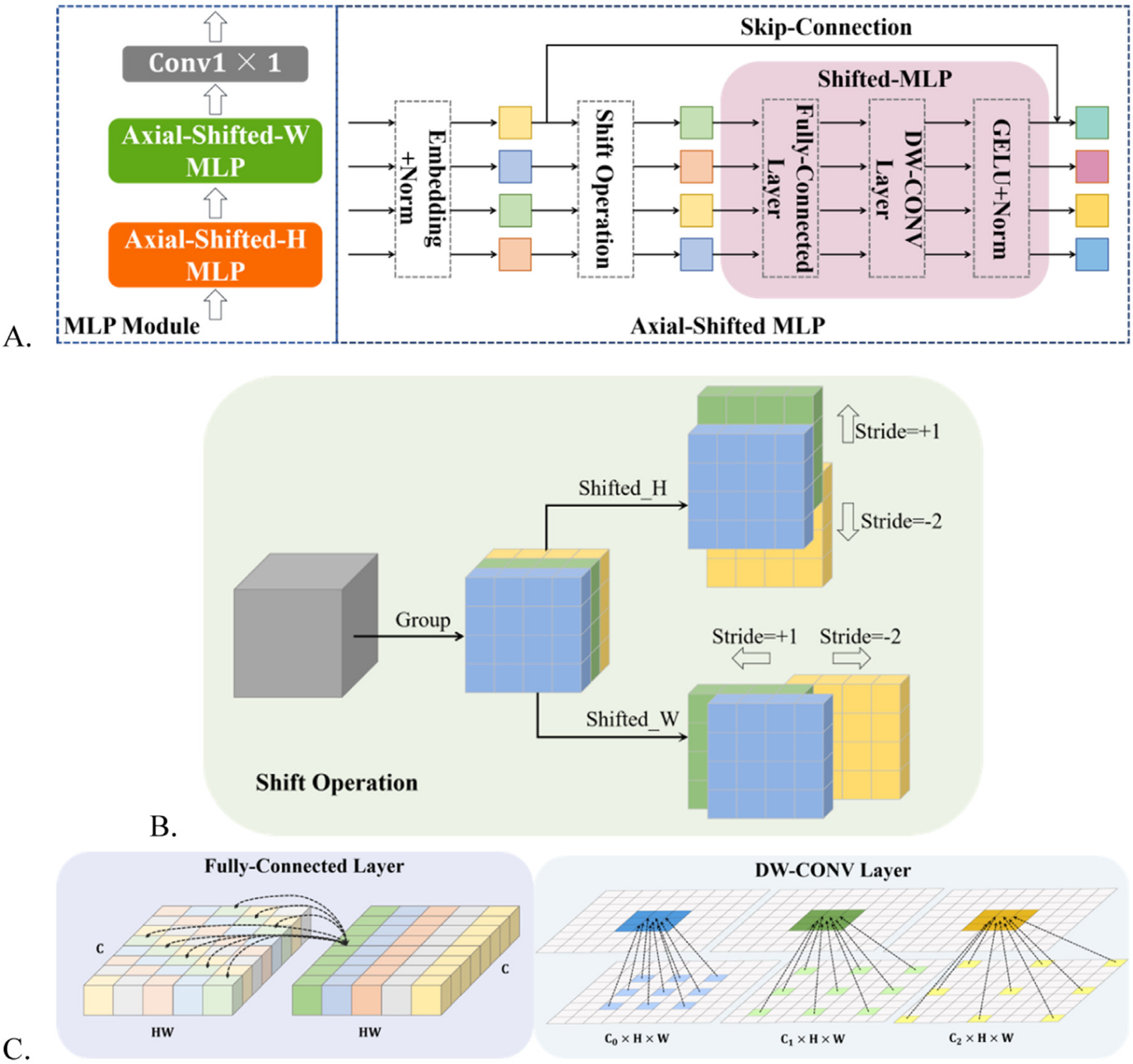

Fig. 3. Overview of the MLP-based module. A) Sub-modules and process of the MLP-based module (left), as well as the assembly of Axial-Shifted MLP (right). B) Examples of the Shift Operation. C) Examples to show the function of Fully-Connected layer and DW-CONV layer.

In summary, the computation in the MLP module can be defined as:

$$Y_H = f_H(X_{in});\ \ Y_W = f_W(Y_H) \tag{5}$$

$$Y_{out} = Conv_{1\times 1}(Y_W) \tag{6}$$

where $X_{in} \in R^{C_{in}\times H\times W}$ and $Y_{out} \in R^{C_{out}\times\frac{H}{2}\times\frac{W}{2}}$ represent input and output feature maps respectively. $f_H(.)$ and $f_W(.)$ refer to the Axial-Shifted-H MLP and Axial-Shifted-W MLP, respectively. $Conv_{1\times 1}$ aims at creating the same channel for the output of MLP-based module and CNN-based module. We assign C_{out} as 32, 64, 128, and 256 in the four stages of parallel branch individually.

3.2. Feature-interaction module

In order to refine the features extracted from the MLPs branch and CNNs branch, we designed a novel feature-interaction module (FIM) to guide context affinity between the two branches. Inspired by the TransFuse [49], which entails a BiFusion module to combine the features from CNN and transformer, we adopted a cross-spatial-wise and cross-channel-wise attention to mold the FIM. A schematic illustration of our proposed FIM is presented in Fig. 4. The following paragraphs elaborate the cross-spatial-wise attention and cross-channel-wise attention.

3.2.1. Cross-spatial-wise attention

Adding the spatial information of feature in MLPs branch into the feature in CNNs branch can instrumentally strengthen its long-range semantic representation, with the construction of cross-spatial-wise attention displayed in Fig. 4a). We define feature map of the MLPs branch and CNNs branch as $Xm \in R^{C\times H\times W}$ and $Xc \in R^{C\times H\times W}$, respectively. Next, the Xc passes through the weight layers $W^q(x)$ and $W^v(x)$ to obtain $SQ \in R^{C\times H\times W}$ and $SV \in R^{C\times H\times W}$ separately, while $SK \in R^{C\times H\times W}$ is obtained after Xm input into the $W^k(x)$. The $SQ \in R^{C\times H\times W}$, $SK \in R^{C\times H\times W}$ and $SV \in R^{C\times H\times W}$ are directly reshaped into $SQ \in R^{C\times N}$, $SK \in R^{N\times C}$ and $SV \in R^{C\times N}$ (N denotes $H\times W$). Subsequently, we use matrix multiplication between $SQ \in R^{C\times N}$ and $SK \in R^{N\times C}$, and then feed multiplication result into the softmax layer to obtain the cross-spatial-wise attention map $S \in R^{N\times N}$, which is calculated as follows:

$$s_{j,i} = \frac{exp\left(sk_i \cdot sq_j\right)}{\sum_{i=1}^{N} exp\left(sk_i \cdot sq_j\right)} \tag{7}$$

where $s_{j,i}$ reflects the effect of the j^{th} spatial position in SQ on the i^{th} spatial position in SK. Further, we implement another matrix

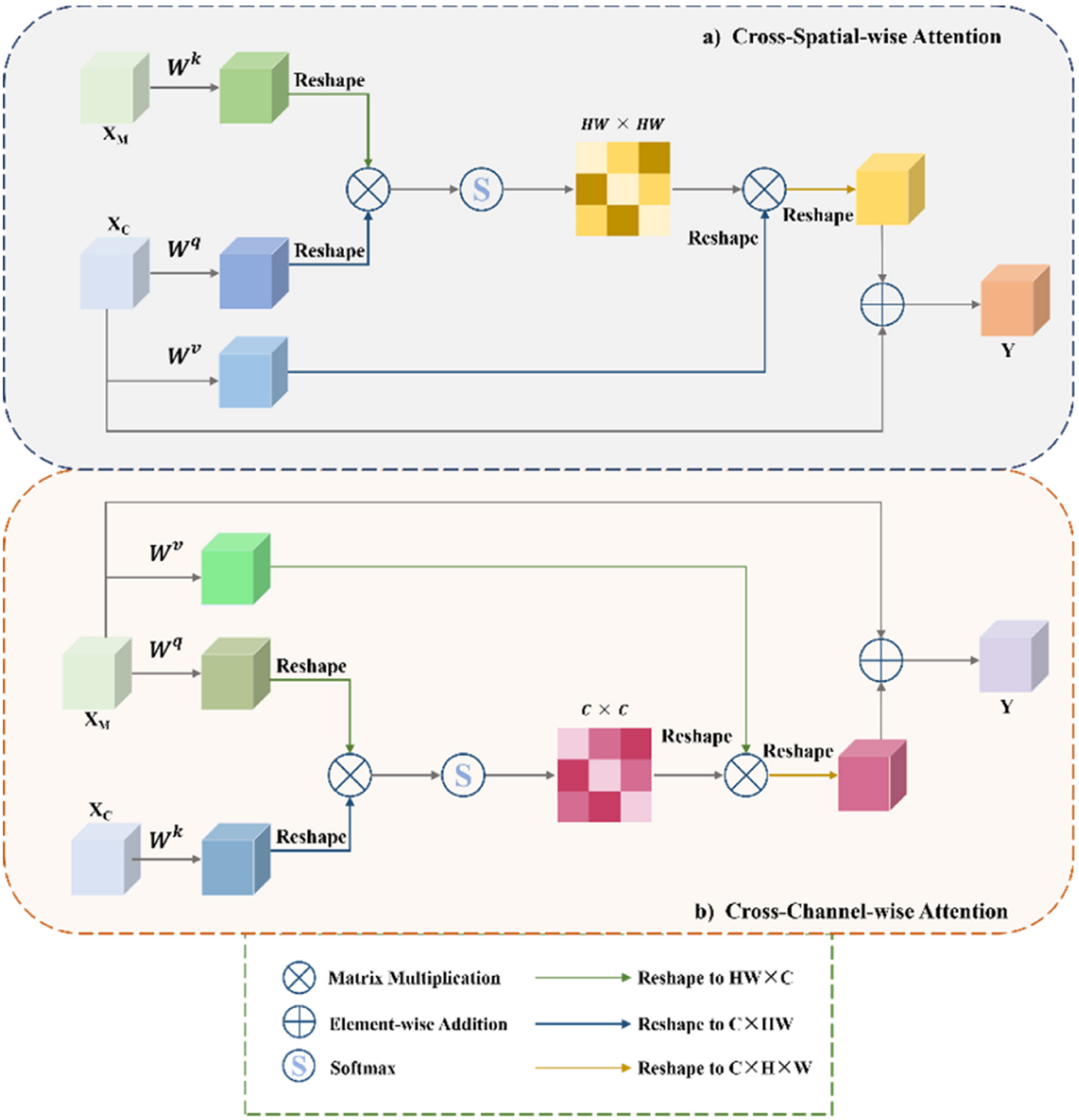

Fig. 4. Overview of the proposed FIM. a) Cross-Spatial-wise Attention; b) Cross-Channel-wise Attention. W^q, W^k and W^v stand for three 1 × 1 convolution layers with different weights.

multiplication between $SV \in R^{C\times N}$ and $S \in R^{N\times N}$, and reshape the multiplication output to $R^{C\times H\times W}$. The final enhanced CNNs feature maps $Yc \in R^{C\times H\times W}$ can be calculated as below:

$$Yc_{j,i} = \omega_C \sum_{i=1}^{N} \left(s_{j,i} \cdot sv_i\right) + Xc_j \tag{8}$$

where ω_C is a learnable weight and initialized to zero. The output Yc has a global cross feature semantic representation compared with Xc, which means that Xc selectively aggregates the spatial context information of Xm, improving the spatial interaction between features.

3.2.2. Cross-channel-wise attention

Dual attention network (DAN) exploits the interdependencies between channel maps to enhance the feature representation of specific semantics [50]. Based on this premise, the cross-channel-wise attention is deployed to build channel interaction between MLPs features and CNNs features, introducing specific semantic responses to MLPs features from CNNs features. The details of cross-channel-wise attention are illustrated in Fig. 4b), in which $CQ \in R^{C\times H\times W}$ and $CV \in R^{C\times H\times W}$ are acquired after $Xm \in R^{C\times H\times W}$ goes through weight layers $W^q(x)$ and $W^v(x)$, and $Xc \in R^{C\times H\times W}$ are fed into $W^k(x)$ to generate the $CK \in R^{C\times H\times W}$. Subsequently, we reshape $CQ \in R^{C\times H\times W}$, $CK \in R^{C\times H\times W}$, and $CV \in R^{C\times H\times W}$ to $CQ \in R^{N\times C}$, $CK \in R^{C\times N}$, and $CV \in R^{N\times C}$, and conduct the matrix multiplication between CQ and CK. The multiplication output then passes through softmax layer to generate the cross-channel-wise attention map $C \in R^{C\times C}$, where it performs matrix multiplication with $CV \in R^{N\times C}$ with the result reshaped to obtain the final output $Ym \in R^{C\times H\times W}$. The above process can be described as:

$$c_{j,i} = \frac{exp\left(ck_i \cdot cq_j\right)}{\sum_{i=1}^{N} exp\left(ck_i \cdot cq_j\right)} \tag{9}$$

$$Ym_{j,i} = \omega_M \sum_{i=1}^{N} \left(c_{j,i} \cdot cv_i\right) + Xm_j \tag{10}$$

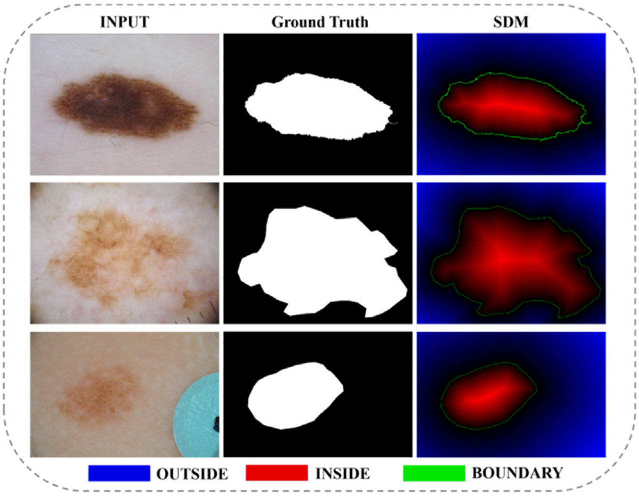

Fig. 5. SDMs of the skin lesions in the ISIC 2018 dataset. The red and blue areas represent the closest distances from the inside pixels and the outside pixels to the boundaries, respectively. Deeper color corresponds to higher magnitude of distance. The boundary is contoured by green color.

where ω_M is also a learnable weight, initializing to zero. The output Ym acts as the weighted sum of all cross-channel features, which is to establish the specific semantic association between Xm and Xc.

To summarize, our FIM can dynamically gather the spatial information of MLPs features into CNNs features, and can also merge the channel information of CNNs features into MLPs features, for the purpose of refining semantic representation of the network.

3.3. Multi-task loss function

For enhancing the learning ability of the network, we design the multi-task strategy, in which the output of network is composed of binary segmentation map (BSM) and signed distance map (SDM). The loss function is formed by two parts as elaborated below.

3.3.1. BCE-dice loss

After the BSM passes through the sigmoid activation**,** we conduct a hybrid loss L_{BSM} including binary cross-entropy (BCE) loss and dice loss [51] to perform the BSM prediction task, which is formulated as follows:

$$L_{bce} = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(p_i) + (1-y_i) log(1-P_i)] \tag{11}$$

$$L_{Dice} = 1 - \frac{2\sum_i^N p_i y_i + \varepsilon}{\sum_i^N p_i + \sum_i^N y_i + \varepsilon} \tag{12}$$

$$L_{BSM} = L_{bce} + L_{Dice} \tag{13}$$

where N denotes the total number of output pixels and $y_i \in \{0, 1\}$ refers to the ground truth of the i^{th} pixel. In addition, $p_i \in [0, 1]$ is the predicted probability of the pixel, which belongs to the foreground class. We set ε as the 10^{-5} to guarantee the stability of data.

3.3.2. SDM loss

Fig. 5 visualizes the SDMs of skin lesions. We can see that SDMs indicate whether the pixels are inside or outside the boundaries of the skin lesions. Compared with BSMs, SDMs contain global geometric information, which helps the network to learn shape and position of skin lesions. Subsequently, we explain how to calculate the SDM loss. Firstly, we define the image and corresponding ground truth as X and Y, and we convert the Y to the SDM S according to the function $T(x)$ as follows:

$$T(x) = \begin{cases} -\underset{y\in B}{inf}\|x-y\|_2 \ , & x \in Y_{in} \\ 0 \ , & x \in B \\ +\underset{y\in B}{inf}\|x-y\|_2 \ , & x \in Y_{out} \end{cases} \tag{14}$$

where B, Y_{in}, and Y_{out} denote the skin lesion boundary, skin lesion regions, and background regions, respectively. Particularly, the $\pm\underset{y\in B}{inf}\|x-y\|_2$ represents the closest distance from background pixel and skin lesion pixel to the boundary, while zero distance indicates the boundary pixel. We further normalize the outside distance value to be in the range $(0, +1]$ and the inside distance value to be in the range $[-1, 0)$, namely the range of value of the S is $[-1, 1]$. Besides, we employ the tanh activation to the SDM prediction and then obtain the $Y_{SDM} \in [-1, 1]$. Next, we calculate the SDM loss according to the following formula:

$$L_{SDM} = \frac{1}{N}\sum_{i=1}^{N}||Y_{SDM}^{i} - S^{i}||^2 \tag{15}$$

where N represents the total pixel numbers. The SDM prediction task promotes the network to learn the position and geometry information, so as to contract the over-segmentation and under-segmentation errors. When co-training SDM and BSM, the final loss L is defined as follows:

$$L = L_{BSM} + \omega L_{SDM} \tag{16}$$

where the ω is a trade-off value between L_{BSM} and L_{SDM}, which is set at 0.35 in our experiments.

4. Experiments and results

4.1. Materials

In this work, we used four public skin lesion datasets to assess the segmentation performance of CFF-Net, i.e., ISIC 2016 [25], ISIC 2017 [24], ISIC 2018 [22,23], and PH2 [26] datasets. We also evaluated CFF-Net on three public datasets (i.e., CVC–ColonDB [52], BUSI [53], and Nasopharyngeal Carcinoma) to demonstrate its applicability to other segmentation tasks.

4.1.1. Skin lesion datasets

ISIC 2016 comprises 1279 images, of which 720, 180, and 379 images were used to train, validate, and test, respectively. ISIC 2017 contains 2750 images, which were divided into 2000, 150, and 600 for training, validation, and testing, respectively. ISIC 2018 contains 2594 images, which were divided into 1816, 260, and 518 for training, validation, and testing, respectively. PH2 is a small dataset with only 200 dermoscopic images. For more reliable assessment of performance, we used 120, 30, and 50 images for training, validation, and testing, respectively. For a more credible evaluation, we adopted 5-fold cross-validation in comparison experiments for ISIC 2016, ISIC 2017, and ISIC 2018 datasets, and 3-fold cross-validation for the PH2 dataset. In addition, we performed 5-fold cross-validation on the ablation experiments of ISIC 2016 and ISIC 2018, and 3-fold cross-validation on that of PH2.

All the dermoscopic images were resized to the resolution of 256 × 256 and normalized to the standard normal distribution. To further improve the generalizability of our proposed model, we employed online random horizontal-vertical flipping and 90° rotating augmentation for all the images.

4.1.2. Other public datasets

4.1.2.1. CVC–ColonDB. Colorectal cancer (CRC) is the fourth most common cause of cancer-related mortality worldwide [54]. Polyps in the intestinal mucosa are believed to be the precursor lesions for CRC [55]. Thus, early detection and diagnosis of polyps are important. CVC–ColonDB comprises 380 colonoscopy images at the resolution 500 × 574 pixels gained from 15 short video sequences with ground truths (GTs). We divided the dataset into training set (200), validation set (50), and testing set (130) and used 5-fold cross-validation for comparison experiments.

4.1.2.2. BUSI. Breast cancer is one of the most fatal cancers in women. Early diagnosis and treatment of breast cancer can help improve the prognosis. An effective means for automatic segmentation of breast lesions from the ultrasound images can facilitate the diagnosis of breast cancer. BUSI is a repository of 780 images from 600 women, which include 133 normal cases, 437 benign cases, and 210 malignant cases. Since clinicians usually focus on the breast lesion area, we removed the normal cases and adopted 3-fold comparison experiments to assess the segmentation performance of each model.

4.1.2.3. NPC. Nasopharyngeal carcinoma (NPC) has a high incidence rate in Southeast Asia, North Africa, and the Middle East [56]. Chemotherapy and radiotherapy are the standard treatment options for NPC. During radiotherapy, the primary NPC tumor is included in the irradiation field. Therefore, accurate contouring of the primary NPC tumors from medical images is a key step for treatment planning. The public NPC dataset used in this study was from MICCAI 2019 StructSeg Challenge, which contains 50 CT scans. A total of 642 slices with primary NPC tumors areas were extracted along the *Z* axis. The comparison experiments were based on 5-fold cross-validation.

4.2. Evaluation metrics

Metrics intend to evaluate the similarity between the ground truth and the predicted results, thus reflecting the performance of the network. In this work, we selected four kinds of metrics to appraise the performance including the Dice similarity coefficient (DSC), Jaccard Index (JA), Recall (*Re*), and Precision (Pre). These metrics are described as follows:

$$DSC = \frac{2 \times TP}{FN + 2 \times TP + FP} \tag{17}$$

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

$$Re = \frac{TP}{TP + FN} \tag{19}$$

$$Pre = \frac{TP}{TP + FP} \tag{20}$$

where *TP, FP*, and *FN* stand for true-positive, false-positive, and false-negative, respectively. The values of *JA, DSC, Re*, and *Pre* range between 0 and 1. Specially, the value of the above metrics stays positively with the network segmentation performance.

4.3. Implementation details

The proposed CFF-Net was implemented in the Pytorch platform and NVIDIA 2080Ti. Adam was selected as our optimizer, in which β_1 and β_2 were set as 0.9 and 0.999, respectively. Additionally, the initial learning rate was 0.0001 with a weight decay of 0.0003, and we employed cosine annealing learning rate scheduler, where the minimum learning rate was 0.00001. Moreover, we chose 8 and 300 as the batch size and epochs for training. For fair comparison, all models were trained with the same hyperparameter. Of note, the BSM activated by sigmoid was binarized with a threshold of 0.5 as the final prediction.

4.4. Comparison experiments

To study the performance of the proposed CFF-Net, **comparison** experiments were conducted on ISIC 2018, ISIC 2017, ISIC 2016, and PH2 datasets using the following nine state-of-the-art models: U-Net (2015) [4], U-Net++ (2018) [5], Att-UNet (2018) [57], CE-Net (2019) [58], CPF-Net (2020) [7], MS RED (2022) [8], FAT-Net (2022) [14], Swin-UNet (2021) [12], and UNeXt (2022) [17]. In particular, Swin-Unet is a model based on the pure Swin Transformer [43] for multi-organ segmentation task, and FAT-Net is a model for skin lesion segmentation, in which the encoder is constructed by CNNs and transformers in parallel. UNeXt is the initial proposed convolutional MLP-based network for medical image segmentation [17]. In other words, we not only compared CFF-Net with the traditional advanced models based on CNNs, but also with those introducing transformers and MLPs, whose experimental results are presented in Table 1. Among these models, CFF-Net showed the best figures in nine of the twelve metrics among the four public skin lesion datasets.

4.4.1. Quantitative evaluation

As observed in Table 1, regarding the four public datasets, our proposed CFF-Net adopts the interactive dual branch encoder synthesized by MLPs and CNNs and leverages the geometry regularization, demonstrating better segmentation metrics on JA (81.86%), DSC (89.78%), and *Re* (88.14%) in the ISIC 2018 dataset; JA (80.21%), DSC (88.69%), and Pre (90.65%) in the ISIC 2017 dataset; JA (85.38%), DSC (92.00%), and Pre (92.58%) in the ISIC 2016 dataset, and JA (89.71%), DSC (94.52%), and *Re* (94.32%) in the PH2 dataset. Thus, the segmentation performance of our method outperformed that of the variant models based on CNNs [4,5,7,8,57,58] and the models with transformers and MLPs [12,14,17].

4.4.2. Qualitative evaluation

Fig. 6 visualizes the comparative results of ISIC 2018, ISIC 2017, ISIC 2016, and PH2. The images in the first and second rows of Fig. 6 show the segmentation results of different models in the presence of interference factors (e.g., hair and color calibration chart). The comparison methods can easily mis-classify the normal tissue as lesion areas in the absence of an excellent feature extraction ability. Because of the dual branch encoder with FIM, which simultaneously captures both local and global context information for enhancing the feature representations of network, CFF-Net can reduce false-positives and more accurately identify the skin lesion areas, as shown in the last column of Fig. 6. The images in the fifth and sixth rows of Fig. 6 present the segmentation results of different methods for cases in which the edges of skin lesions are unclear and the contrast between the skin lesions and the background is low. It is apparent that CFF-Net provides better segmentation performance in this situation, as the auxiliary loss on SDM prediction improves the boundary awareness of the network. In general, the segmentation results (see the last column in Fig. 6) indicate that CFF-Net can effectively recover finer details and mitigate the issues of under-segmentation and over-segmentation compared with the other state-of-the-art models. This is because the dual branch encoder with FIM enables the network to pay more attention to the semantic information of skin lesions, and the multi-task loss helps the network to more accurately locate the boundaries of skin lesions.

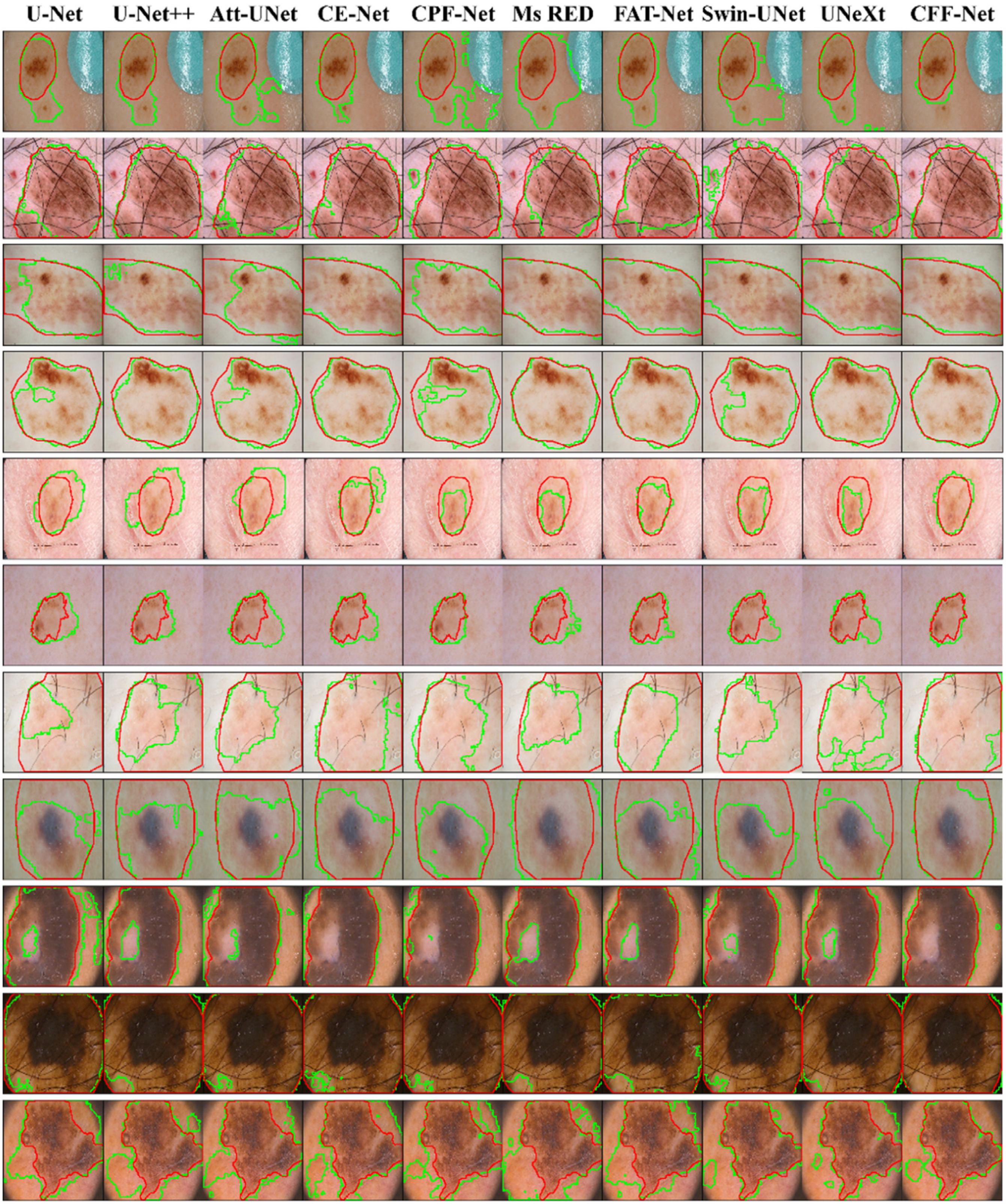

Fig. 6. Visualization of comparison results in the ISIC 2018, ISIC 2017, ISIC 2016, and PH2 datasets. The regions enclosed by red and green denote the ground truth (GT) and the segmentation prediction, respectively.

Table 1
Performance comparison of the skin lesion segmentation in ISIC 2018, ISIC 2017, ISIC 2016 and PH2 datasets (mean± SD). The most favorable results are emphasized in bold.

Dataset	Model	JA (%)	DSC (%)	*Re* (%)	Pre (%)
ISIC 2018	U-Net [4]	79.71 ± 0.76	88.43 ± 0.33	85.89 ± 0.69	89.46 ± 0.52
	U-Net++ [5]	81.24 ± 0.26	89.43 ± 0.26	87.94 ± 0.50	89.45 ± 0.41
	Att-UNet [57]	79.52 ± 0.75	88.32 ± 0.33	87.11 ± 0.63	87.73 ± 0.57
	CE-Net [58]	81.32 ± 0.68	89.45 ± 0.30	87.60 ± 0.59	89.18 ± 0.49
	CPF-Net [7]	80.45 ± 0.82	88.86 ± 0.36	87.57 ± 0.65	88.44 ± 0.61
	MS RED [8]	80.92 ± 0.81	89.15 ± 0.37	87.36 ± 0.60	89.14 ± 0.70
	FAT-Net [14]	80.92 ± 0.73	89.18 ± 0.32	87.88 ± 0.56	88.52 ± 0.54
	UNeXt [17]	79.73 ± 0.80	88.43 ± 0.35	86.78 ± 0.58	88.47 ± 0.68
	Swin-UNet [12]	79.86 ± 0.91	88.46 ± 0.41	87.43 ± 0.67	88.24 ± 0.73
	CFF-Net	81.86 ± 0.69	89.78 ± 0.30	88.14 ± 0.49	89.40 ± 0.53
ISIC 2017	U-Net [4]	78.03 ± 0.81	87.34 ± 0.38	84.52 ± 0.90	89.07 ± 0.46
	U-Net++ [5]	77.34 ± 0.84	86.90 ± 0.39	83.64 ± 0.97	88.56 ± 0.46
	Att-UNet [57]	77.54 ± 0.91	86.99 ± 0.44	84.99 ± 0.99	87.29 ± 0.62
	CE-Net [58]	79.34 ± 0.88	88.15 ± 0.41	84.99 ± 0.86	89.91 ± 0.49
	CPF-Net [7]	79.12 ± 0.94	87.99 ± 0.44	86.83 ± 0.80	87.67 ± 0.64
	MS RED [8]	79.38 ± 0.86	88.17 ± 0.86	86.89 ± 0.76	87.07 ± 0.64
	FAT-Net [14]	79.12 ± 0.77	88.05 ± 0.36	86.54 ± 0.80	87.77 ± 0.53
	UNeXt [17]	78.62 ± 0.88	87.69 ± 0.42	85.47 ± 0.89	88.51 ± 0.56
	Swin-UNet [12]	72.28 ± 1.00	83.50 ± 0.52	81.02 ± 1.25	84.27 ± 0.71
	CFF-Net	80.21 ± 0.86	88.69 ± 0.40	85.66 ± 0.90	90.65 ± 0.38
ISIC 2016	U-Net [4]	82.58 ± 0.50	90.29 ± 0.20	87.72 ± 0.63	91.29 ± 0.13
	U-Net++ [5]	83.25 ± 0.41	90.72 ± 0.16	88.48 ± 0.47	91.80 ± 0.17
	Att-UNet [57]	83.43 ± 0.41	90.83 ± 0.16	88.42 ± 0.51	92.12 ± 0.12
	CE-Net [58]	85.01 ± 0.30	91.80 ± 0.11	90.32 ± 0.39	91.84 ± 0.13
	CPF-Net [7]	84.24 ± 0.32	91.34 ± 0.12	89.78 ± 0.42	91.34 ± 0.15
	MS RED [8]	84.43 ± 0.36	91.43 ± 0.14	89.53 ± 0.38	91.60 ± 0.20
	FAT-Net [14]	84.49 ± 0.30	91.49 ± 0.11	90.41 ± 0.36	91.11 ± 0.19
	UNeXt [17]	82.59 ± 0.45	90.31 ± 0.18	84.18 ± 1.18	90.49 ± 0.22
	Swin-UNet [12]	84.14 ± 0.34	91.27 ± 0.13	89.45 ± 0.42	91.39 ± 0.15
	CFF-Net	85.38 ± 0.33	92.00 ± 0.12	90.36 ± 0.34	92.58 ± 0.17
PH2	U-Net [4]	84.18 ± 0.36	91.30 ± 0.12	89.86 ± 0.56	90.30 ± 0.46
	U-Net++ [5]	84.76 ± 0.56	91.57 ± 0.21	90.88 ± 0.56	90.45 ± 0.44
	Att-UNet [57]	82.70 ± 0.67	90.29 ± 0.28	88.45 ± 0.57	91.35 ± 0.78
	CE-Net [58]	89.29 ± 0.16	94.29 ± 0.53	94.31 ± 0.08	93.21 ± 0.29
	CPF-Net [7]	88.03 ± 0.30	93.54 ± 0.10	94.09 ± 0.08	91.81 ± 0.42
	MS RED [8]	88.18 ± 0.25	93.65 ± 0.08	93.52 ± 0.18	86.38 ± 0.85
	FAT-Net [14]	87.29 ± 0.20	93.15 ± 0.06	92.51 ± 0.24	92.85 ± 0.33
	UNeXt [17]	85.69 ± 0.78	92.01 ± 0.35	92.87 ± 0.16	90.12 ± 0.96
	Swin-UNet [12]	87.08 ± 1.05	92.69 ± 0.55	94.29 ± 0.06	91.42 ± 1.29
	CFF-Net	89.71 ± 0.18	94.52 ± 0.06	94.32 ± 0.08	91.55 ± 0.21

4.4.3. Attention visualization

The visualization results of different models, depicting the regions of interest, are shown in Fig. 7. Compared with the other state-of-the-art models, our model shows significantly improved ability to locate skin lesion areas and filter the background noise. For cases with low contrast or unclear boundary, our model can highlight skin lesions successfully and retain more local details, since the design of dual branch encoder enables the network to learn both local and global spatial information. In addition, the introduction of auxiliary prediction task for SDM promotes the capture of global spatial relationship between foreground and background pixels, which further helps the model to pay more attention to the boundaries of the lesion. Moreover, our model can mitigate the interference of artifacts (e.g., hair and color calibration) due to the representation enhancement by FIM.

4.5. Ablation experiments

To investigate the effect of each component on our proposed method, we performed ablation experiments by sequentially removing SDM prediction-head, FIM, and MLPs branch from the CFF-Net.

4.5.1. Quantitative evaluation

Table 2 presents the segmentation metrics of CFF-Net and the models after sequential detachment of each module on the three public datasets. Especially, JA is the more important metric than the others in the skin lesion segmentation challenge; therefore, the drop value of JA is a critical index to measure the contribution of each part. Data presented in Table 2 shows that CFF-Net performs best in ten of the twelve metrics in the three public skin lesion datasets. The removal of SDM prediction-head brings a JA drop of 0.48%, 0.52%, and 0.71% for the ISIC 2018, ISIC 2016, and PH2 datasets, respectively. Additionally, the JA decrease is 0.73% for the ISIC 2018, 0.68% for the ISIC 2016, and 0.65% for the PH2 after further separating the FIM. Finally, withdrawing the MLPs branch from the network encoder resulted in a JA drop of 0.21% for ISIC 2018, 0.51% for ISIC 2016, and 2.13% for PH2. In general, the addition of these three components gradually raised the segmentation accuracy of the network.

4.5.2. Qualitative evaluation

We carried out qualitative analysis to increase the interpretability of three proposed components. First, we visualized the results of the models before and after removing the SDM prediction-head and those of the four advanced models [14,17,57,58], in order to explain the superiority of introducing geometric information into the network, as elaborated in Fig. 8. Below, we provide the definition of correct-segmentation area, under-segmentation area, and over-segmentation area of Fig. 8. Given the binarization prediction $P \in \{0, 1\}^{H \times W}$ and GT $G \in \{0, 1\}^{H \times W}$, the correct-segmentation area $C \in \{0, 1\}^{H \times W}$ can be equated as $P \times G$. Furthermore, the under-segmentation area $U \in \{0, 1\}^{H \times W}$ and over-segmentation area $O \in \{0, 1\}^{H \times W}$ can be equated as $G - C$ and $P - C$, re-

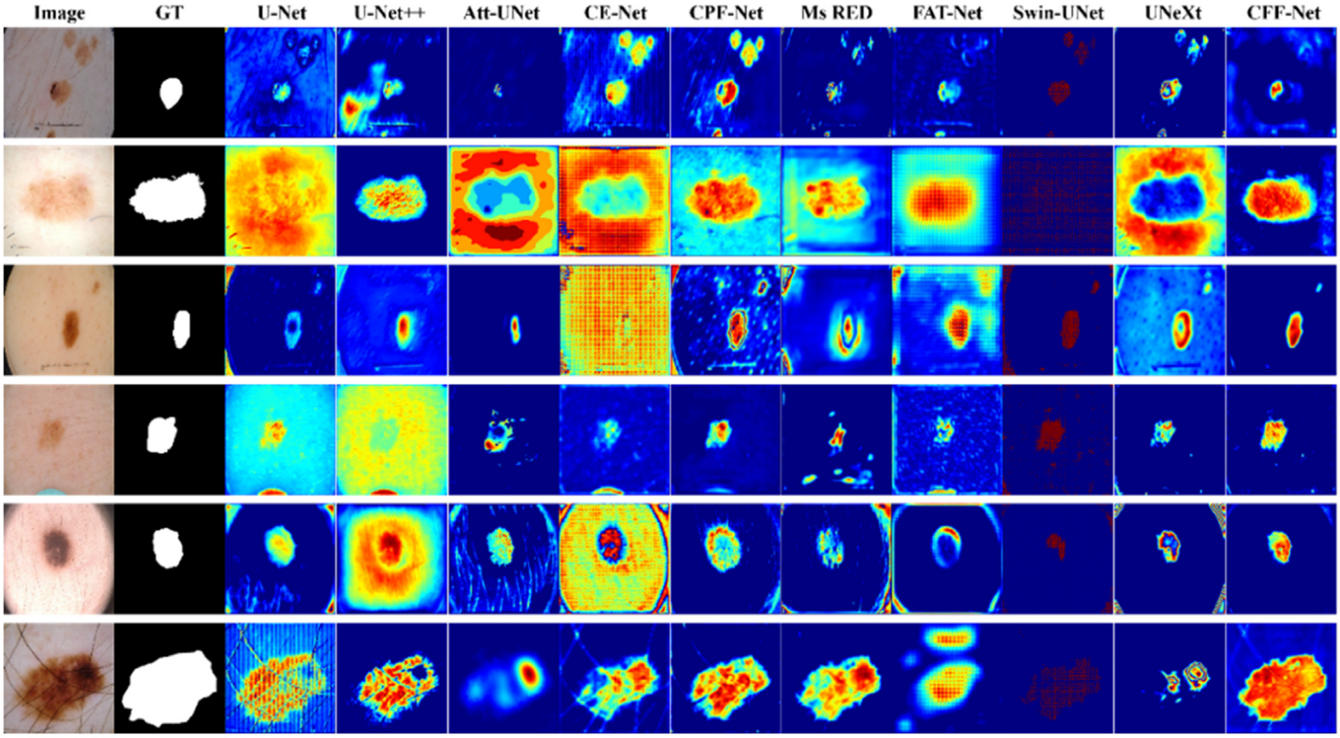

Fig. 7. The attention maps obtained by the last layer decoder from different models. Note that, warmer color indicates higher attention scores.

Table 2
Ablation studies of different components in the ISIC 2018, ISIC 2016, and PH2 (mean± SD). Note that the ↓ is the drop value of JA compared with the previous model. The best results are depicted in bold.

Model	JA drop (%)	JA (%)	DSC (%)	*Re* (%)	Pre (%)
		ISIC 2018			
CFF-Net	–	81.86 ± 0.69	89.78 ± 0.30	88.14 ± 0.49	89.40 ± 0.53
-SDM	↓0.48	81.38 ± 0.77	89.46 ± 0.33	87.15 ± 0.65	90.11 ± 0.54
-FIM	↓0.73	80.65 ± 0.61	89.07 ± 0.26	87.52 ± 0.57	88.55 ± 0.46
-MLPs	↓0.21	80.44 ± 0.67	88.92 ± 0.29	87.47 ± 0.57	88.99 ± 0.50
		ISIC 2016			
CFF-Net	–	85.38 ± 0.33	92.00 ± 0.12	90.36 ± 0.34	92.58 ± 0.17
-SDM	↓0.52	84.86 ± 0.39	91.68 ± 0.15	89.80 ± 0.48	92.24 ± 0.11
-FIM	↓0.68	84.18 ± 0.38	91.23 ± 0.14	89.28 ± 0.42	91.92 ± 0.17
-MLPs	↓0.51	83.67 ± 0.42	90.97 ± 0.16	89.38 ± 0.49	91.44 ± 0.18
		PH2			
CFF-Net	–	89.71 ± 0.18	94.52 ± 0.06	94.32 ± 0.08	91.55 ± 0.21
-SDM	↓0.71	89.00 ± 0.32	94.08 ± 0.11	95.02 ± 0.56	87.99 ± 0.48
-FIM	↓0.65	88.35 ± 0.15	93.77 ± 0.48	93.64 ± 0.12	89.16 ± 0.25
-MLPs	↓2.13	86.22 ± 0.91	92.27 ± 0.43	93.49 ± 0.17	89.91 ± 1.01

spectively. It is worth noting that the smaller regions of yellow and green (larger regions of white) in Fig. 8 represent the better segmentation performance of network. By contrasting with the models with only the binary segmentation map (BSM) supervision, auxiliary learning of SDM deploys position and shape awareness, producing clearer contours and decreasing the false-positives. Overall, the combination training of SDM and BSM retains great skin lesion details and aligns well with the GT.

Besides, we utilized attention maps to exhibit features extracted by the last layer of CNNs encoder, MLPs encoder as well as the CNNs features after spatial information exchange, manifesting the feasibility of feature-interaction. Several examples, which include three skin lesion types, namely melanocytic nevus (nv), melanoma (mel), and benign keratosis (bkl), are visualized by attention maps in Fig. 9. As observed in Fig. 9(a), CNNs encoder can only highlight the local skin lesion regions, and cannot filter out the background noise. Fig. 9(b) shows that MLPs encoder can significantly improve the accuracy in detecting the whole skin lesion regions from a global receptive field, due to its powerful ability in establishing spatial long-range dependencies. Although MLPs encoder has advantages in capturing the global context information, it still cannot generate skin lesion targets with smooth boundaries and reduce false-positives. Judged from Fig. 9(c), our spatial interaction skill promotes CNNs features to incorporate into global semantic information from MLPs features, enhancing the contrast between the lesion and the surrounding background, and suppressing the irrelevant noise. Specially, our proposed encoder included CNNs branch and MLPs branch can extract finer features from the dermoscopy images attributed to the categories of nv, mel, and bkl, which enhances the clinical practicality of CFF-Net.

4.6. Cross validation experiments on ISIC 2018 and PH2

In order to estimate the segmentation effectiveness of CFF-Net with inconsistent distribution of training dataset and test

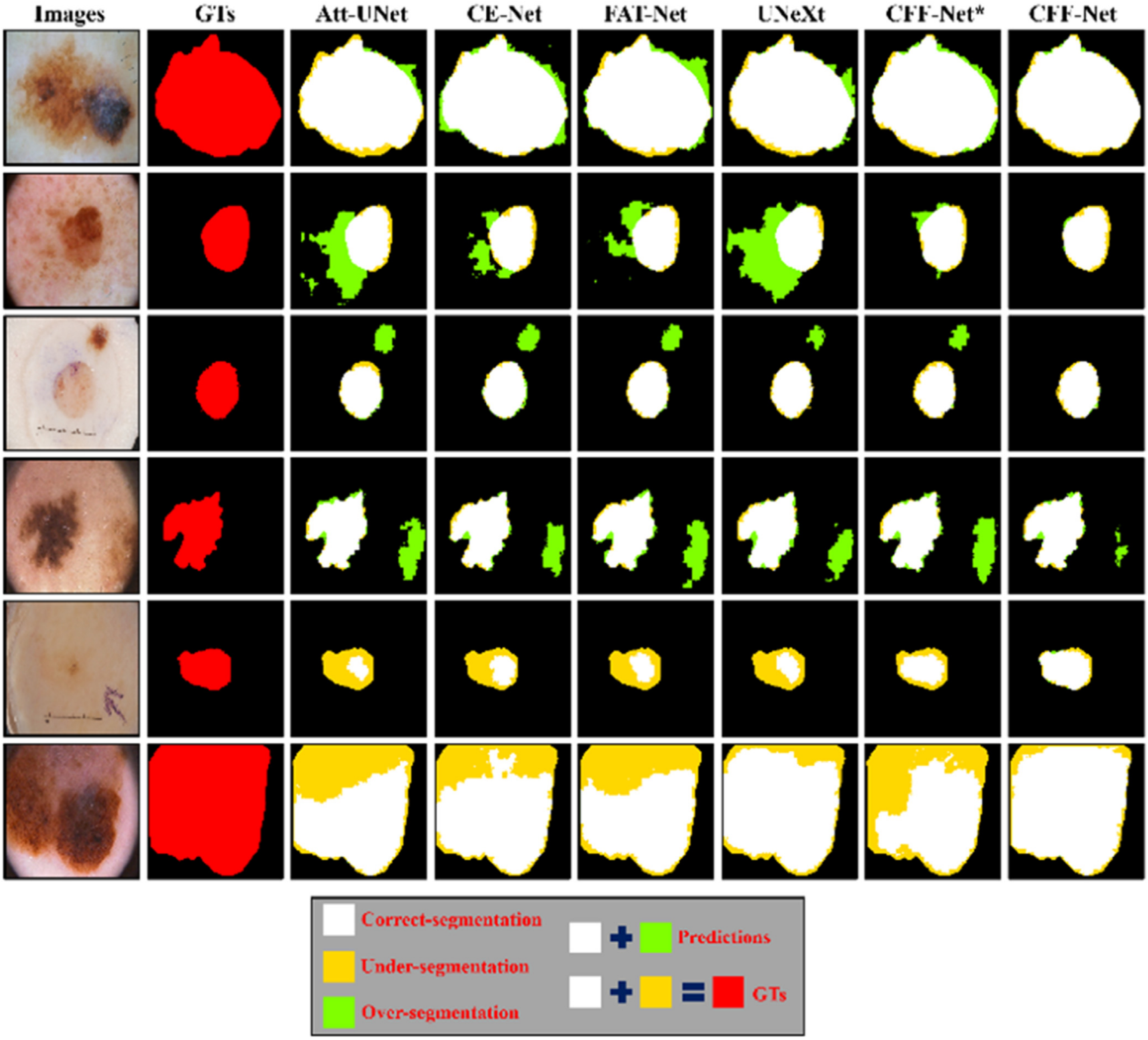

Fig. 8. Visualization of ablation results on ISIC 2018, ISIC 2016, and PH2 datasets. 'CFF-Net*' refers to the model after removing SDM prediction-head from the CFF-Net. White, yellow, and green represent the correct-segmentation areas, under-segmentation areas, and over-segmentation areas, respectively. Moreover, red represents the GTs, which are composed of correct-segmentation areas and under-segmentation areas.

dataset, we implemented cross validation between ISIC 2018 and PH2 (Table 3). For this experiment, all models to be tested were trained on ISIC 2018 and PH2 using 5-fold cross-validation and 3-fold cross-validation, respectively. Next, we selected 200 data from PH2 to evaluate the models obtained from ISIC 2018, and 500 data from the ISIC 2018 to evaluate the models obtained from PH2. As shown in Table 3, CFF-Net exhibits promising performance with a JA (54.31%) and a DSC (68.53%) in the mode of 'ISIC 2018 → PH2', as well as a JA (59.92%) and a DSC (69.29%) in the mode of 'PH2 → ISIC 2018′. This indicates that our proposed method is more robust than the others. Moreover, we generated boxplots to present descriptive statistics of JA and DSC on the two different modes (Fig. 10). The boxplots show that the divergence between cases of cross validation is reduced by our method, which helps yield highest mean value and median value in both JA and DSC metrics.

4.7. Comparison experiments on other segmentation tasks

In order to further demonstrate the universality and robustness of the proposed network, we conducted comparison experiments on three public datasets (i.e., CVC–ColonDB, BUSI, and NPC).

4.7.1. CVC–ColonDB

As illustrated in Table 4, the quantitative results show that the CFF-Net possesses the highest metrics of JA (83.99%), DSC (91.07%), and Pre (89.84%) compared with the six state-of-the-art models in CVC–ColonDB, implying that the segmentation predictions produced by our method are closer to the ground truth. Qualitative results are presented in Fig. 11, and it is evident that most models encountered some challenges in pointing out the vivid boundaries and reconstructing the details of target areas when the lesions in the colonoscopy images exhibited irregular shapes and blurry

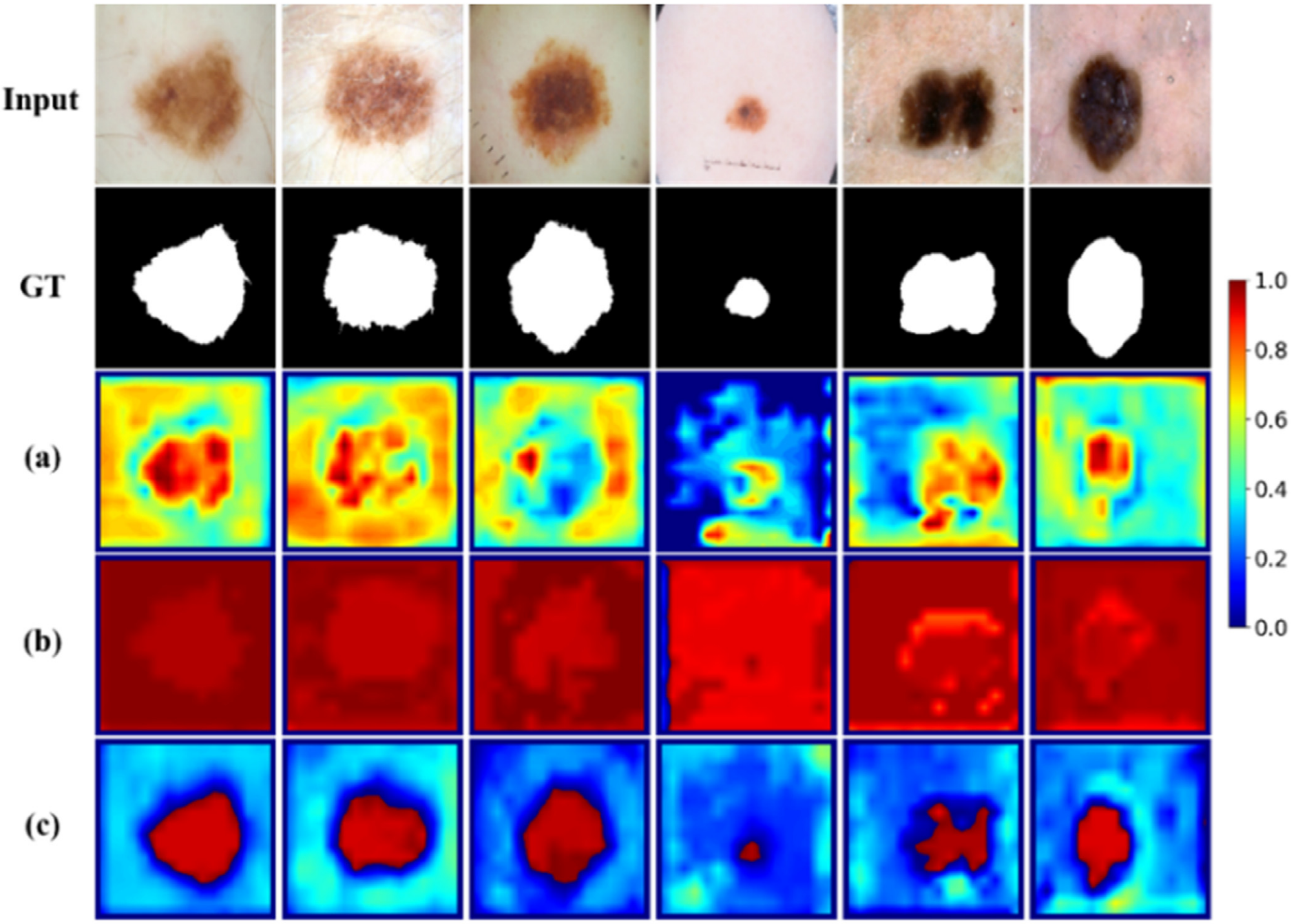

Fig. 9. Visualization of the feature maps generated by the last layer of CNNs encoder and MLPs encoder. (a) the CNNs feature maps; (b) the MLPs feature maps; (c) the enhanced CNNs feature maps (i.e., skip-connections). Note that warmer color refers to higher attention scores. The first two columns, the middle columns, and the last two columns of samples belong to the nv, mel, and bkl categories, respectively.

Table 3
Cross validation between ISIC 2018 and PH2 on different models (mean± SD). 'ISIC 2018 → PH2' means that the models are trained on the ISIC 2018 and tested on the PH2, while 'PH2 → ISIC 2018′ represents the opposite steps. The best results are indicated in bold.

	ISIC 2018 → PH2		PH2 → ISIC 2018	
Model	JA (%)	DSC (%)	JA (%)	DSC (%)
U-Net [4]	24.47 ± 5.47	34.14 ± 7.74	50.15 ± 10.23	60.01 ± 10.48
U-Net++ [5]	18.40 ± 2.78	29.32 ± 2.78	54.49 ± 8.92	64.96 ± 8.57
Att-UNet [57]	37.91 ± 2.98	52.67 ± 3.49	55.12 ± 8.78	65.61 ± 8.41
CE-Net [58]	49.28 ± 2.70	64.45 ± 2.44	56.99 ± 9.80	66.49 ± 9.66
CPF-Net [7]	49.52 ± 0.88	65.69 ± 0.79	58.39 ± 9.45	67.80 ± 9.06
MS RED [8]	50.21 ± 3.43	64.76 ± 2.91	59.48 ± 9.52	68.81 ± 9.22
FAT-Net [14]	38.51 ± 3.05	53.38 ± 3.18	53.70 ± 9.70	63.66 ± 9.69
UNeXt [17]	28.58 ± 1.11	43.38 ± 1.76	58.78 ± 8.35	69.06 ± 7.71
Swin-UNet [12]	47.92 ± 3.78	62.28 ± 3.72	55.71 ± 9.52	65.60 ± 9.30
CFF-Net	54.31 ± 3.23	68.53 ± 2.60	59.92 ± 9.38	69.29 ± 8.96

Table 4
Comparison of our model with the field-leading methods using the CVC–ColonDB dataset (mean± SD). The best results are indicated in bold.

Model	JA (%)	DSC (%)	*Re* (%)	Pre (%)
U-Net [4]	79.76 ± 0.83	88.43 ± 0.38	87.02 ± 0.58	88.78 ± 0.67
U-Net++ [5]	77.59 ± 1.64	86.66 ± 1.01	83.86 ± 1.46	83.79 ± 1.09
Att-UNet [57]	79.39 ± 1.51	87.85 ± 0.97	84.85 ± 1.46	87.03 ± 0.79
CE-Net [58]	76.70 ± 1.22	86.34 ± 0.58	86.32 ± 0.77	86.40 ± 1.11
Swin-Unet [12]	51.15 ± 1.77	66.66 ± 1.35	68.65 ± 1.97	76.70 ± 1.22
PraNet [59]	83.43 ± 0.64	90.74 ± 0.27	90.52 ± 0.30	89.70 ± 0.56
CFF-Net	83.99 ± 0.65	91.07 ± 0.28	90.30 ± 0.30	89.84 ± 0.83

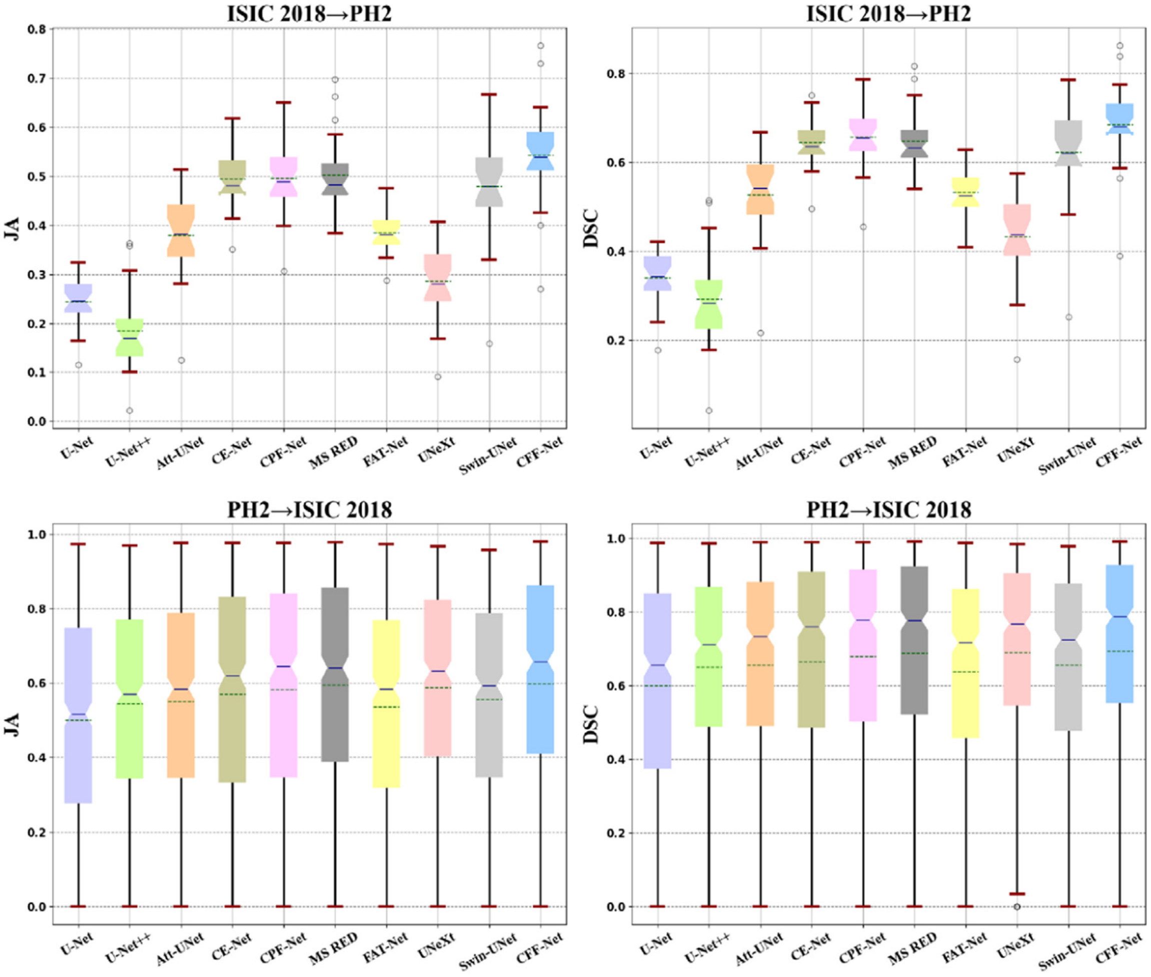

Fig. 10. Boxplots of the cross validation experiments on ISIC 2018 and PH2. The green line '–' and blue line '-' denote the mean value and median value, respectively.

Table 5
Comparison of our model with the field-leading models using the BUSI dataset (mean± SD). The best results are indicated in bold.

Model	JA	DSC	*Re*	Pre
U-Net [4]	62.20 ± 2.03	75.74 ± 1.20	72.13 ± 1.69	79.45 ± 1.78
U-Net++ [5]	61.26 ± 2.16	74.91 ± 1.38	71.27 ± 1.77	74.90 ± 2.16
Att-UNet [57]	63.58 ± 1.97	76.80 ± 1.21	72.15 ± 1.82	83.24 ± 1.39
CE-Net [58]	63.59 ± 2.40	76.57 ± 1.55	76.17 ± 1.50	76.42 ± 2.33
UNeXt [17]	52.24 ± 2.74	67.01 ± 2.26	61.86 ± 3.00	75.45 ± 1.98
MedT [13]	50.09 ± 2.34	65.26 ± 2.16	64.68 ± 2.29	64.93 ± 2.71
CFF-Net	64.15 ± 2.37	77.06 ± 2.37	75.80 ± 1.40	77.53 ± 2.05

boundaries, as well as low contrast with the background. On the contrary, our method can enable more accurate localization of targets and generate more continuous segmentation maps, since the FIM dynamically aggregates local and global contexts from the dual branch encoder to generate richer feature representations. Moreover, the introduction of SDM supervision further calibrates the segmentation results.

4.7.2. BUSI

Quantitative verification of the results in the BUSI are presented in Table 5. the highest JA (64.15%) and DSC (77.06%) were obtained with the CFF-Net. The experiments demonstrate that the breast lesion predictions by our algorithm show a better match with the ground truth. Furthermore, the qualitative results illustrated in Fig. 12 revealed that most models tended to mis-detect the edges of the breast lesions, along with neglecting the details of breast lesions. Yet, with the assistance of geometric constraint and long-range dependencies comprised of global-spatial-dependencies and global-channel-dependencies, our proposed approach showed better performance in handling the breast lesions with blurry edges and irregular shape.

4.7.3. NPC

The challenge in this segmentation task lies in the low contrast between the primary NPC tumors and the surrounding normal tissues. We compared our model with six advanced models in this dataset, and the experimental results are reported in Table 6. Our method shows the highest JA (64.93%), DSC (77.03%), and Pre (79.76%), further demonstrating that CFF-Net is an effective and robust method. Some visual examples of these comparison experiments are presented in Fig. 13. It can be observed that our proposed model reduces false-positive predictions and produces sharp

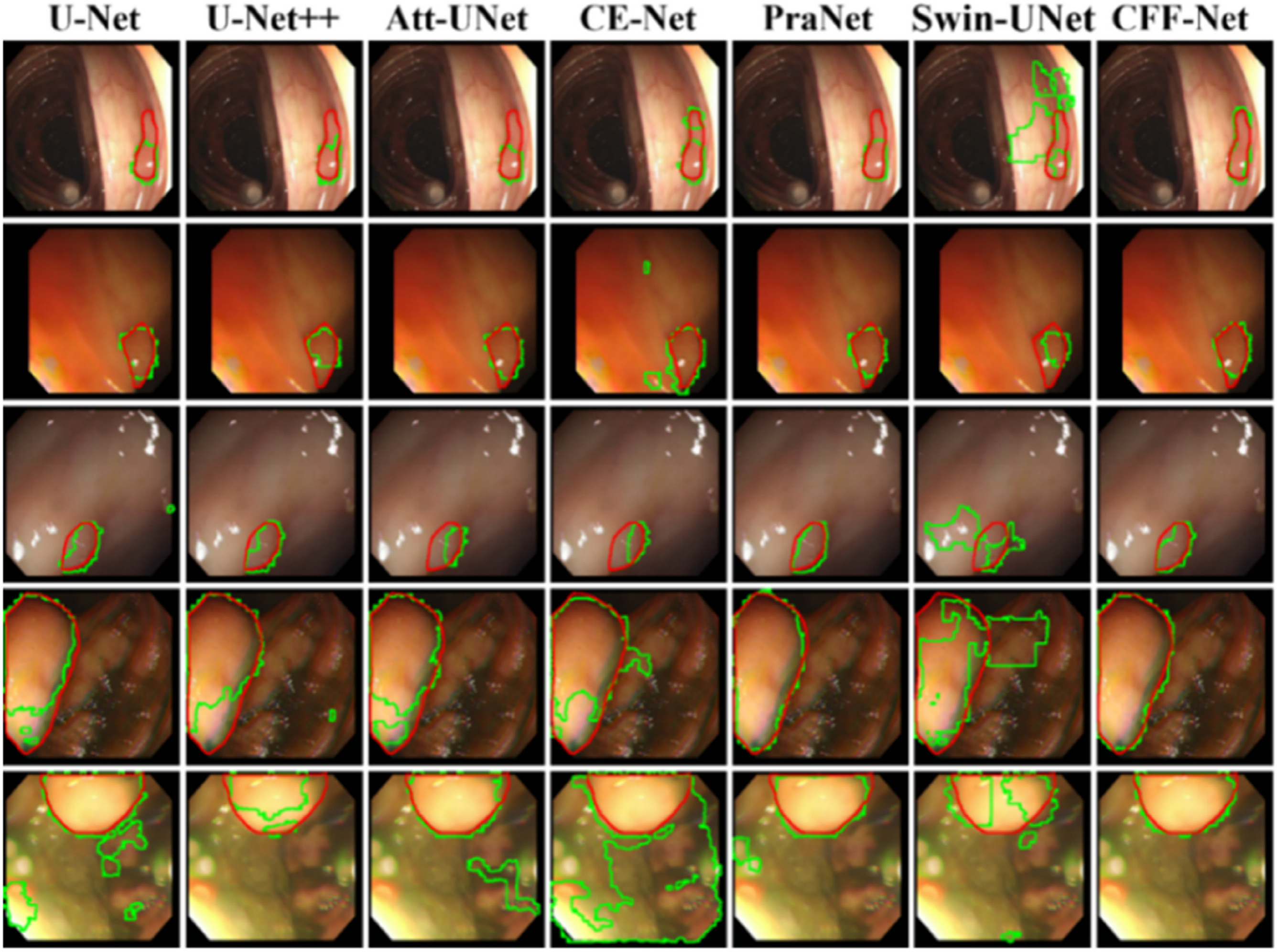

Fig. 11. Visualization of comparative results of different models in the CVC–ColonDB dataset. The red and green contours mark the boundaries of ground truth and segmentation predictions, respectively.

Table 6
Comparison of our model with the field-leading methods in the NPC dataset (mean± SD). The best results are indicated in bold.

Model	JA	DSC	*Re*	Pre
U-Net [4]	62.99 ± 2.96	75.66 ± 2.47	79.49 ± 2.86	78.13 ± 4.09
U-Net++ [5]	60.57 ± 3.13	73.68 ± 2.66	77.93 ± 3.50	76.14 ± 4.51
Att-UNet [57]	60.26 ± 3.23	73.34 ± 2.86	78.67 ± 3.51	75.33 ± 4.71
CE-Net [58]	60.78 ± 3.08	73.85 ± 2.65	78.95 ± 3.46	75.68 ± 4.15
UNeXt [17]	64.29 ± 2.89	76.70 ± 2.35	80.36 ± 3.00	78.94 ± 3.39
FAT-Net [14]	60.06 ± 3.15	73.23 ± 2.78	79.42 ± 3.43	74.04 ± 4.47
CFF-Net	64.93 ± 3.10	77.03 ± 2.65	79.92 ± 3.32	79.76 ± 3.32

boundaries of primary NPC tumors, i.e., the segmentation results of CFF-Net are more aligned with the ground truth than the compared models. This is because the CFF-Net employs a strong encoder which fuses the local features and global contextual information extracted from the CNNs and MLPs, leading to strong representation capability. Moreover, the auxiliary prediction task is introduced to guide the learning of global geometric information, yielding more complete primary NPC tumor areas and suppressing the irrelevant noise.

4.8. Computational complexity

The overall training time and inference time of all models in ISIC 2018 are shown in Table 7. In this experiment, it takes approximately 2.9 h to train the CFF-Net (i.e., per epoch). In addition, CFF-Net costs less than 1 s to generate one mask of skin lesion (i.e., 6 s for 518 dermoscopy images). Compared with other models, CFF-Net has moderate training and inference time. The parameters (Parms) and floating-pointing operations per second (FLOPs) are presented in Table 7. In contrast with the models based on pure CNNs (i.e., U-Net, U-Net++, Att-UNet, CE-Net, CPF-Net, and MS RED), CFF-Net has comparable Parms and FLOPs. Moreover, our model is more lightweight than both FAT-Net and Swin-UNet equipped with transformers. Although the UNeXt has advantages in terms of Parms and FLOPs, its segmentation accuracy is lower than that of the other models compared. Overall, our proposed method achieves a good balance between segmentation performance and computational complexity.

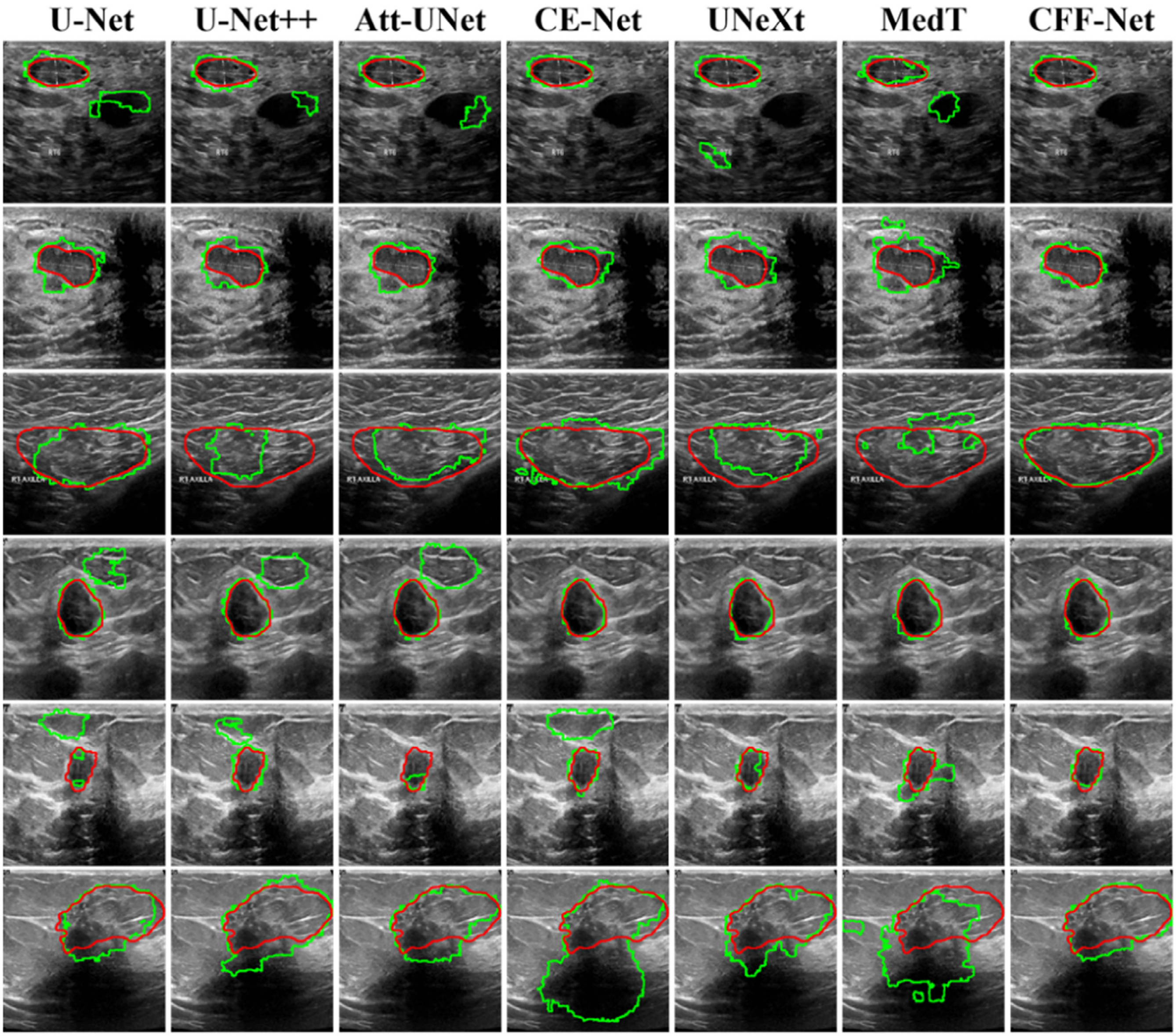

Fig. 12. Visualization of the comparative results of different models in the BUSI dataset. The red and green contours indicate the boundaries of ground truth and segmentation predictions. Particularly, the first three rows show benign predictions while the last three rows show malignant predictions.

Additionally, the Parms of CFF-Net after removing SDM prediction-head (S-p), MLPs branch (M-b), and FIM are illustrated in Table 8. The parameters of CFF-Net before and after removing S-p are almost the same, because we only used a simple convolution layer to output the prediction of SDM. Moreover, our FIM costs about 0.11 M parameters, which accounts for only 1.1% for the whole segmentation model. Although the S-p and FIM are lightweight, they both bring significant improvement in JA metrics of ISIC 2018, ISIC 2016, and PH2. Moreover, MLPs branch, which has acceptable Parms (1.75 M), also enhances the segmentation performance of network. To summarize, our proposed modules have comparable parameters and all improve the segmentation accuracy of the network.

5. Discussion and limitations

Developing an accurate and effective automatic segmentation method is important for computer-aided analysis and diagnosis of medical imaging [60,61]. In this work, we designed a novel model for automatic skin lesion segmentation in dermoscopy images which may facilitate the primary detection and early treatment of melanoma. All in all, CFF-Net is equipped with three proposed components: the dual branch encoder composed of MLPs and CNNs, extracting rich contextual information; the FIM to enhance the feature representations of MLPs branch and CNNs branch; SDM prediction-head to promote shape and position awareness, highlighting the boundaries of segmentation. Regarding the experimental results, CFF-Net transcended the nine state-of-the-art models, and had the best JA (81.86%), DSC (89.78%), and *Re* (88.14%) in ISIC 2018, the best JA (80.21%), DSC (88.69%), and Pre (90.65%) in ISIC 2017, the best JA (85.38%), DSC (92.00%), and Pre (92.58%) in ISIC 2016, as well as the best JA (89.71%), DSC (94.52%), and *Re* (94.32%) in PH2. The other metrics (i.e., Pre and *Re*) are presented in Table 1. Furthermore, some visual examples of comparison results are displayed in Fig. 6, showing that CFF-Net performs precise boundary segmentation while restoring more details.

The ablation results showed the contributions of dual branch encoder, FIM, and SDM prediction-head to the segmentation

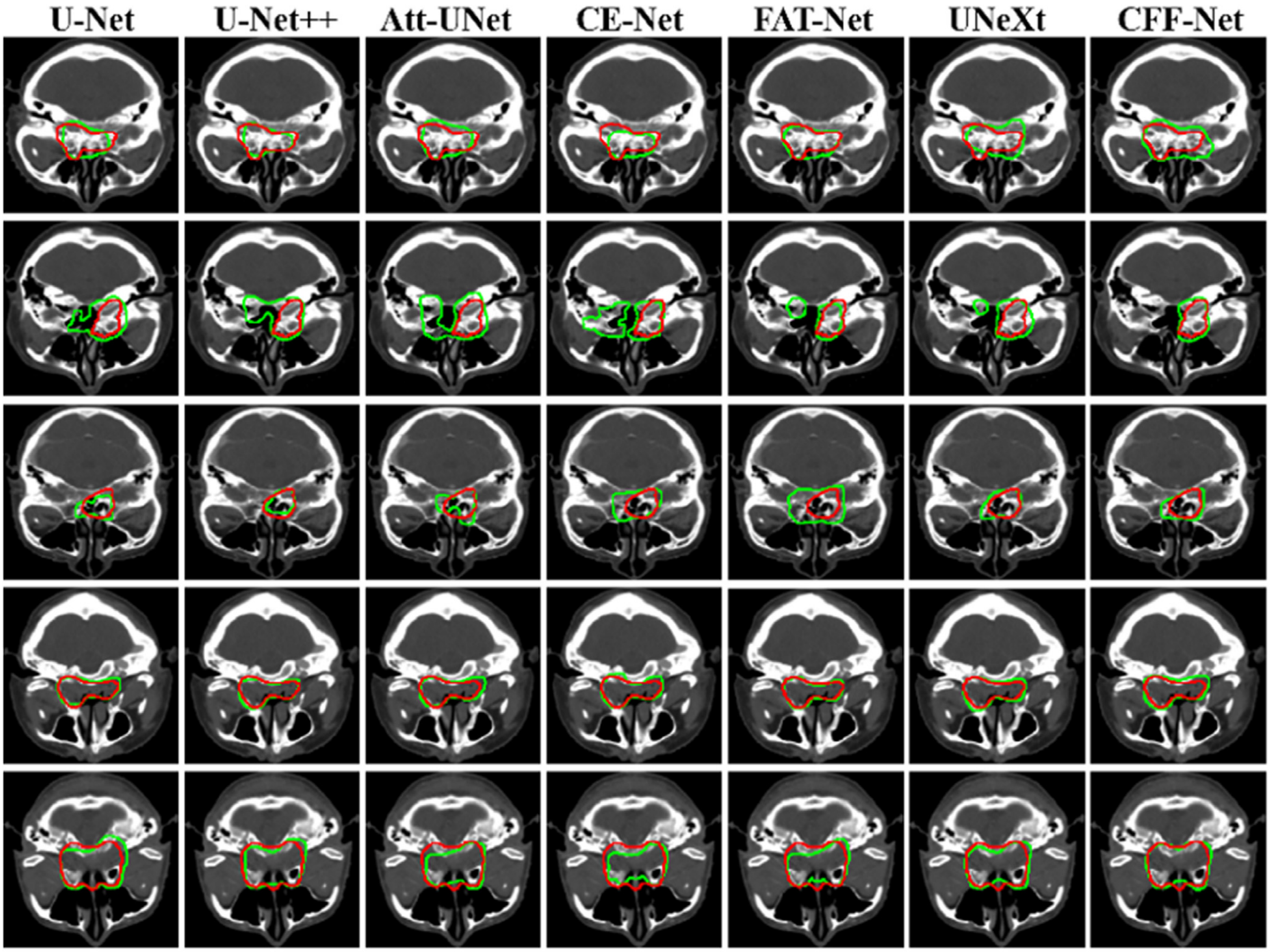

Fig. 13. Visualization of the comparative results of different models in the NPC dataset. The red and green contours indicate the boundaries of ground truth and segmentation predictions.

Table 7
Computational complexity and segmentation performance of different methods in ISIC 2018.

Method	FLOPs (G)	Parms (M)	Training Time (h)	Inference Time (s)	JA (%)
U-Net [4]	20.18	9.85	2.0	7	79.71 ± 0.76
U-Net++ [5]	35.02	9.34	4.6	9	81.24 ± 0.26
Att-UNet [57]	16.71	8.73	2.7	6	79.52 ± 0.75
CE-Net [58]	8.89	29.00	1.6	4	81.32 ± 0.68
CPF-Net [7]	8.03	30.65	2.3	5	80.45 ± 0.82
MS RED [8]	10.55	4.71	4.8	11	80.92 ± 0.81
FAT-Net [14]	42.83	29.61	3.2	6	80.92 ± 0.73
UneXt [17]	1.02	0.25	0.8	3	79.73 ± 0.80
Swin-Unet [12]	5.86	27.12	3.0	5	79.86 ± 0.91
CFF-Net	14.70	9.71	2.9	6	81.86 ± 0.69

Table 8
Parameters and performance of different modules in ISIC 2018, ISIC 2016, and PH2.

Method	Parms (M)	JA (%) /ISIC 2018	JA (%) /ISIC 2016	JA (%) /PH2
CFF-Net	9.71	81.86 ± 0.69	85.38 ± 0.33	89.71 ± 0.18
CFF-Net - S-p	9.71	81.38 ± 0.77	84.86 ± 0.39	89.00 ± 0.32
CFF-Net - S-p - FIM	9.60	80.65 ± 0.61	84.18 ± 0.38	88.35 ± 0.15
CFF-Net - S-p - FIM - M-b	7.85	80.44 ± 0.67	83.67 ± 0.42	86.22 ± 0.91

performance, and the results are reported in Table 2. As observed in Fig. 8, co-training of SDM and BSM enabled the depiction of the geometrical contour of the target. Recently, the MLPs consisting of fully connected layers and non-linear activation functions have received extensive attention [62,63], owing to the simple and effective channel-mixing and spatial-mixing operations. To the best of our knowledge, CFF-Net is the first model to aggregate the MLPs and CNNs in parallel for medical image segmentation, expressing both local and global receptive fields. Moreover, we integrated the spatial information into MLPs from CNNs, and the channel information into CNNs from MLPs, enabling more significant semantic representations for further decoding. Furthermore, we adopted visual attention maps to explain the MLPs features and CNNs features before and after improvement, as shown in Fig. 9.

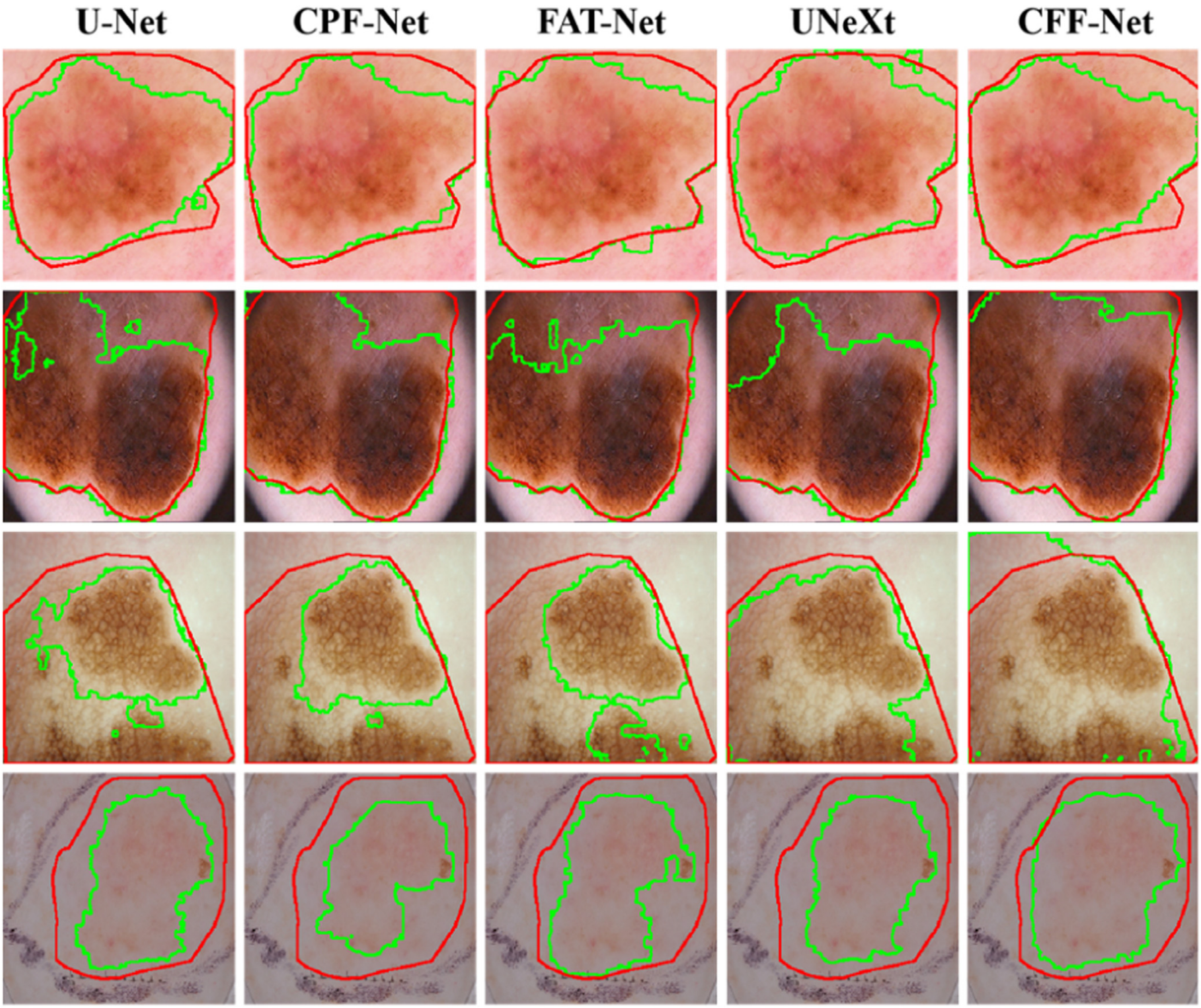

Fig. 14. Visual comparisons of failure cases with different methods. The regions enclosed by red and green denote GT and segmentation prediction, respectively.

Finally, we conducted cross-validation between ISIC 2018 and PH2, and the results are presented in Table 3. Besides, the boxplots of JA and DSC (Fig. 10) illustrated that the median value of CFF-Net was highest while the deviation was smallest for both metrics. Furthermore, we executed other three segmentation tasks (i.e., CVC−ColonDB, BUSI, and NPC) using our model, and the results are reported in Tables 4, 5, and 6, respectively. For CVC−ColonDB, JA and DSC increased by 4.23% and 2.64% using our model compared to U-Net, while for BUSI, the CFF-Net surpassed the U-Net by 1.95% on JA and 1.32% on DSC. In addition, For NPC, the JA and DSC improved by 1.94% and 1.37%. Figs. 11, 12, and 13 show the contour maps of predictions and ground truth, indicating the superiority of our model for recovering details, decreasing false-positives, and highlighting the small targets with ambiguous boundaries. These experiments demonstrated the effectiveness and robustness of our model in providing precise and reliable automatic segmentation.

Although CFF-Net showed advanced performance in four skin lesion datasets, it achieved limited segmentation accuracy for individual cases. Fig. 14 shows some cases of false segmentation by our proposed CFF-Net and other four models. Similar to the models based on CNNs [4,7], transformers [14], and MLPs [17], our model also showed poor segmentation performance in cases where the skin lesions were oversized or carried extremely low contrast to the background. Although our model cannot fully identify skin lesion areas, the segmentation results of our method were closest to the ground truth compared with other models. In the future, we will explore a novel combination mode between MLPs and CNNs to enhance the feature extraction ability and improve the segmentation performance.

6. Conclusion

In this study, we propose a novel architecture CFF-Net which is equipped with the encoder composed of MLPs and CNNs in parallel together with integrating feature-interaction module (FIM) to allow the features of different branches to dynamically exchange spatial information and channel information. This design enhanced the capability of representation learning. Besides, we leverage an auxiliary loss on supervising geometric shapes, which enables more accurate contouring of the boundaries of skin lesions. Extensive experiments on four publicly available skin lesion datasets demonstrated that our proposed model possesses leading-edge segmentation performance with comparable computational complexity, yielding prediction maps with more details, clearer boundaries, and less false-positives. Furthermore, we conducted comprehensive ablation experiments to assess the effect of the individual components of the proposed model. Application of CFF-Net for the other three segmentation tasks showed its universality and robustness. In future, our method may possibly be applied for other medical image segmentation tasks, such as segmentation of retinal edema lesion and multiple organs from fetal MRI.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

Acknowledgement

This study was supported by the NSFC (No. 61771347), the Special Project in Key Areas of Artificial Intelligence in Guangdong Universities (No. 2019KZDZX1017), and the Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011576), and the Basic Research and Applied Basic Research Key Project in General Colleges and Universities of Guangdong Province (2021ZDZX1032), and the Guangdong Postgraduate Education Innovation Program Project (YJS-SFJD-21–02, [2022] No.1).

References

[1] A. Jemal, Cancer Statistics, 2017, CA Cancer J. Clin. 67 (2017) 7–30.
[2] P. Mathur, K. Sathishkumar, M. Chaturvedi, P. Das, F.S. Roselind, Cancer Statistics, 2020: report From National Cancer Registry Programme, India, JCO Glob. Oncol. 6 (6) (2020) 1063–1075.
[3] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 3431–3440.
[4] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[5] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: a nested U-Net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, Granada, Spain, 2018, pp. 3–11.
[6] Ran Gu, et al., CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation, IEEE Trans. Med. Imag. 40 (2) (2020) 699–711.
[7] Shuanglang Feng, et al., CPFNet: context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imag. 39 (10) (2020) 3008–3018.
[8] Duwei Dai, et al., Ms RED: a novel multi-scale residual encoding and decoding network for skin lesion segmentation, Med. Image Anal. 75 (2022) 102293.
[9] Ashish Vaswani, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[10] 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 16x16 words: transformers for image recognition at scale, (2020), doi:10.48550/arXiv.2010.11929.
[11] Chen, Jieneng, et al., Transunet: transformers make strong encoders for medical image segmentation, 2021 arXiv preprint arXiv:2102.04306.
[12] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-Unet: unet-like pure transformer for medical image segmentation, 2021 arXiv preprint arXiv:2105.05537.
[13] Jeya Maria Jose Valanarasu, et al., Medical transformer: gated axial-attention for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2021, pp. 36–46, doi:10.1007/978-3-030-87193-2_4.
[14] Huisi Wu, et al., FAT-Net: feature adaptive transformers for automated skin lesion segmentation, Med. Image Anal. 76 (2022) 102327.
[15] Hugo Touvron, et al., Resmlp: feedforward networks for image classification with data-efficient training, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2023) 5314–5321.
[16] Kai Han, et al., A survey on vision transformer, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2022) 87–110.
[17] Valanarasu, Jeya Maria Jose, Vishal M. Patel, UNeXt: MLP-based Rapid Medical Image Segmentation Network, 2022 arXiv preprint arXiv:2203.04967.
[18] Hritam Basak, Rohit Kundu, Ram Sarkar, MFSNet: a multi focus segmentation network for skin lesion segmentation, Pattern Recognit. 128 (2022) 108673.
[19] Rui Gu, Lituan Wang, Lei Zhang, DE-Net: a deep edge network with boundary information for automatic skin lesion segmentation, Neurocomputing 468 (2022) 71–84.
[20] Jeong Joon Park, et al., Deepsdf: learning continuous signed distance functions for shape representation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 165–174.
[21] Xue, Y., Tang, H., Qiao, Z., Gong, G., Yin, Y., Qian, Z., Huang, C., Fan, W., Huang, Shape-aware organ segmentation by predicting signed distance maps, 2019 arXiv preprint arXiv:1912.03849.
[22] Codella, Noel, 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.
[23] Philipp Tschandl, Cliff Rosendahl, Harald Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci Data 5 (1) (2018) 1–9.
[24] Noel CF Codella, 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: IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE, 2018, pp. 168–172.
[25] David Gutman, et al., Skin lesion analysis toward melanoma detection, A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC), 2016 arXiv preprint arXiv:1605.01397.
[26] Zhengzhong Tu, et al., Maxim: multi-axis mlp for image processing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 5769–5780.
[27] Ilya O. Tolstikhin, et al., Mlp-mixer: an all-mlp architecture for vision, Adv. Neural Inf. Process. Syst. 34 (2017) 24261–24272.
[28] Lian, Dongze, et al., As-mlp: an axial shifted mlp architecture for vision, 2021 arXiv preprint arXiv:2107.08391.
[29] Tan Yu, et al., S2-mlp: spatial-shift mlp architecture for vision, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2022, pp. 297–306.
[30] Chen, Shoufa, et al., Cyclemlp: a mlp-like architecture for dense prediction, 2021 arXiv preprint arXiv:2107.10224.
[31] Yan Wang, et al., Deep distance transform for tubular structure segmentation in ct scans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 3833–3842.
[32] Shuailin Li, Chuyu Zhang, Xuming He, Shape-aware semi-supervised 3D semantic segmentation for medical images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 552–561.
[33] Z. Liu, and C. Zhao, Semi-supervised Medical Image Segmentation via Geometry-aware Consistency Training, 2022 arXiv preprint arXiv:2202.06104.
[34] Tran-Dac-Thinh Phan, et al., Skin lesion segmentation by u-net with adaptive skip connection and structural awareness, Appl. Sci. 11 (10) (2021) 4528.
[35] Sarmad Maqsood, Robertas Damaševičius, Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare, Neur. Netw. 160 (2023) 238–258.
[36] Ruiqi Feng, et al., BLA-Net: boundary learning assisted network for skin lesion segmentation, Comput. Methods Programs Biomed. 226 (2022) 107190.
[37] Olusola Oluwakemi Abayomi-Alli, et al., Malignant skin melanoma detection using image augmentation by oversamplingin nonlinear lower-dimensional embedding manifold, Turk. J. Electric. Eng. Comp. Sci. 29 (8) (2021) 2600–2614.
[38] Himanshu K. Gajera, Deepak Ranjan Nayak, Mukesh A. Zaveri, A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features, Biomed. Signal Process. Control. 79 (2023) 104186.
[39] Lequan Yu, et al., Automated melanoma recognition in dermoscopy images via very deep residual networks, IEEE Trans. Med. Imag. 36 (4) (2016) 994–1004.
[40] Seifedine Kadry, et al., Extraction of abnormal skin lesion from dermoscopy image using VGG-SegNet, in: 2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII), IEEE, 2021, pp. 1–5.
[41] Wang, Yaxiong, et al., DONet: dual objective networks for skin lesion segmentation, 2020 arXiv preprint arXiv:2008.08278.
[42] Huisi Wu, et al., Automated skin lesion segmentation via an adaptive dual attention module, IEEE Trans. Med. Imag. 40 (1) (2020) 357–370.
[43] Ruxin Wang, et al., Cascaded context enhancement network for automatic skin lesion segmentation, Expert Syst. Appl. 201 (2022) 117069.
[44] Wenhai Wang, et al., Pyramid vision transformer: a versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 568–578.
[45] Ho, Jonathan, et al., Axial attention in multidimensional transformers, 2019 arXiv preprint arXiv:1912.12180.
[46] Ba, Jimmy Lei, Jamie Ryan Kiros, Geoffrey E. Hinton, Layer normalization, 2016 arXiv preprint arXiv:1607.06450.
[47] Hendrycks, Dan, Kevin Gimpel, Gaussian error linear units (gelus), 2016 arXiv preprint arXiv:1606.08415.
[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[49] Yundong Zhang, Huiye Liu, Qiang Hu, Transfuse: fusing transformers and cnns for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 14–24.
[50] Jun Fu, et al., Dual attention network for scene segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 3146–3154.
[51] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 565–571.
[52] J. Bernal, J. Sánchez, F. Vilarino, Towards automatic polyp detection with a polyp appearance model, Pattern Recogn. 45 (9) (2012) 3166–3182.

[53] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of Breast Ultrasound Images, Data in Brief. 28 (2019) 104863.

[54] Pasqualino Favoriti, et al., Worldwide burden of colorectal cancer: a review, Updates Surg. 68 (2016) 7–11.

[55] Juan José Granados-Romero, et al., Colorectal cancer: a review, Int. J. Res. Med. Sci. 5 (11) (2017) 4667.

[56] Ellen T. Chang, Hans-Olov Adami, The enigmatic epidemiology of nasopharyngeal carcinoma, Cancer Epidemiol. Biomark. Prevent. 15 (10) (2006) 1765–1777.

[57] 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.

[58] Z. Gu, et al., Ce-net: context encoder network for 2d medical image segmentation, IEEE Trans. Med. Imag. 38 (10) (2019) 2281–2292.

[59] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, PraNet: parallel reverse attention network for polyp segmentation, in: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2020, pp. 263–273.

[60] Hao Ma, Yanni Zou, Peter X. Liu, MHSU-Net: a more versatile neural network for medical image segmentation, Comput. Methods Programs Biomed. 208 (2021) 106230.

[61] Jinke Wang, et al., SAR-U-Net: squeeze-and-excitation block and atrous spatial pyramid pooling based residual U-Net for automatic liver segmentation in Computed Tomography, Comput. Methods Programs Biomed. 208 (2021) 106268.

[62] Wenshuo Li, et al., Brain-inspired multilayer perceptron with spiking neurons, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 783–793.

[63] Qibin Hou, et al., Vision permutator: a permutable mlp-like architecture for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2022) 1328–1334.

Chuanbo Qin is an Associate Professor at Wuyi University, Guangdong, China. He received his Ph.D. degree in Pattern Recognition and Intelligent Systems from the Department of Automation Science and Engineering of South China University of Technology in 2015. And he received his B.S and M.S degree both from the Wuyi University, China, in 2004 and 2008, respectively. Since July 2015, he has been working at Department of Intelligence Manufacturing, Wuyi University, Guangdong, China. His main research direction is medical image processing and biometric recognition.

Bin Zheng received his B. Sc. degree in communication engineering from Wuyi University, Guangdong province, China, in 2020. He is currently a M. Sc. candidate of electronic information in the Department of Intelligence Manufacturing at Wuyi University. His main research interests include medical image segmentation and analysis based on the deep learning.

Junying Zeng is an Associate Professor at Wuyi University, Guangdong, China. He received his Ph.D. degree in physical electronics from Beijing University of Posts and Tele-communications, Beijing, China, in 2008. He received the Master degree in physical electronic from Yunnan University, Yunnan, China, in 2005. Since June 2008, he has been working at Department of Intelligence Manufacturing, Wuyi University, Guangdong, China, and his research interests include: image understanding, deep learning and signal processing.

Zhuyuan Chen received his B. Sc. degree in communication engineering from Wuyi University, Guangdong province, China, in 2021. He is currently a M. Sc. candidate of electronic information in the Department of Intelligence Manufacturing at Wuyi University. His main research interests include medical image segmentation and analysis based on the deep learning.

Angelo Genovese received the Ph.D. degree in computer science from the Università degli Studi di Milano, Crema, Italy, in 2014. He has been a postdoctoral Research Fellow in computer science with the Università degli Studi di Milano since 2014. He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. Original results have been published in over 30 papers in international journals, proceedings of international conferences, books, and book chapters. His current research interests include signal and image processing, three-dimensional reconstruction, computational intelligence techn-ologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems.

Vincenzo Piuri is IEEE Fellow, ACM Fellow, and Full Professor at the University of Milan, Italy (since 2000), where he was also Department Chair (2007–2012). He was Associate Professor at Politecnico di Milano, Italy (1992–2000), visiting professor at the University of Texas at Austin, USA (summers 1996–1999), and visiting researcher at George Mason University, USA (summers 2012–2016). He founded a start-up company, Sensuresrl, in the area of intelligent systems for industrial applications (leading it from 2007 until 2010) and was active in industrial research projects with several companies. He received his M.S. and Ph.D. degree in Computer Engineering from Politecnico di Milano, Italy. His main research and industrial application interests are: intelligent systems, computational intelligence, pattern analysis and recognition, machine learning, signal and image processing, biometrics, intelligent measurement systems, industrial applications, distributed processing systems, internet-of-things, cloud computing, fault tolerance, application-specific digital processing architectures, and arithmetic architectures.

Fabio Scotti received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003. He has been an Associate Professor in Computer Science with the Universitá degli Studi di Milano, Crema, Italy, since 2015. Original results have been published in more than 100 papers in international journals, proceedings of international conferences, books, book chapters, and patents. His current research interests include biometric systems, machine learning and computational intelligence, signal and image processing, theory and applications of neural networks, three-dimensional reconstruction, industrial applications, intelligent measurement systems, and high-level system design. He is an Associate Editor with the IEEE Transactions on Human-Machine Systems and Soft Computing (Springer).