Computer Methods and Programs in Biomedicine 242 (2023) 107846

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

MC-DC: An MLP-CNN Based Dual-path Complementary Network for Medical Image Segmentation

Xiaoben Jiang [a], Yu Zhu [a,*], Yatong Liu [a], Nan Wang [a], Lei Yi [b,*]

[a] School of Information Science and Technology, East China University of Science and Technology, Shanghai, 200237, China
[b] Department of Burn, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, 200025, China

ARTICLE INFO

Keywords:
MLP-CNN based dual-path complementary network
dual-path complementarymodule
cross-scale global feature fusion module
cross-scale local feature fusion module
efficient mask feature fusion module
medical image segmentation

ABSTRACT

Background: Fusing the CNN and Transformer in the encoder has recently achieved outstanding performance in medical image segmentation. However, two obvious limitations require addressing: (1) The utilization of Transformer leads to heavy parameters, and its intricate structure demands ample data and resources for training, and (2) most previous research had predominantly focused on enhancing the performance of the feature encoder, with little emphasis placed on the design of the feature decoder.

Methods: To this end, we propose a novel MLP-CNN based dual-path complementary (MC-DC) network for medical image segmentation, which replaces the complex Transformer with a cost-effective Multi-Layer Perceptron (MLP). Specifically, a dual-path complementary (DPC) module is designed to effectively fuse multi-level features from MLP and CNN. To respectively reconstruct global and local information, the dual-path decoder is proposed which is mainly composed of cross-scale global feature fusion (CS-GF) module and cross-scale local feature fusion (CS-LF) module. Moreover, we leverage a simple and efficient segmentation mask feature fusion (SMFF) module to merge the segmentation outcomes generated by the dual-path decoder.

Results: Comprehensive experiments were performed on three typical medical image segmentation tasks. For skin lesions segmentation, our MC-DC network achieved 91.69% Dice and 9.52mm ASSD on the ISIC2018 dataset. In addition, the 91.6% Dice and 94.4% Dice were respectively obtained on the Kvasir-SEG dataset and CVC-ClinicDB dataset for polyp segmentation. Moreover, we also conducted experiments on the private COVID-DS36 dataset for lung lesion segmentation. Our MC-DC has achieved 87.6% [87.1%, 88.1%], and 92.3% [91.8%, 92.7%] on ground-glass opacity, interstitial infiltration, and lung consolidation, respectively.

Conclusions: The experimental results indicate that the proposed MC-DC network exhibits exceptional generalization capability and surpasses other state-of-the-art methods in higher results and lower computational complexity.

1. Introduction

Image segmentation plays a crucial role in medical image analysis, particularly in computer-aided diagnosis and image-guided clinical surgeries [1]. Over the past decade, there has been significant research and development in the field of segmentation, with a focus on developing efficient and robust segmentation methods. U-Net [2] and its variants [3–9] are landmark works that consist of an encoder (downsampling path) and a decoder (upsampling path) in a U-shaped architecture, connected by skip connections. The encoder can model deep semantic information. The decoder then restores the features to the original image size and generates pixel-level segmentation results through upsampling (deconvolution). Moreover, skip connections connect the features from different levels of the encoder with the corresponding levels of the feature map. These integral components enable UNet to capture features at various scales and levels in medical images, such as skin lesions, organs, and cells, facilitating effective segmentation of structures at different scales [10]. In this period, the studies all solely leverage convolution kernels to extract spatial features of images. Despite the notable success of convolutional kernels in detecting local details and edges in medical images, they still exhibit limitations in capturing broader global contextual information.

Recently, Transformers [11] have leveraged self-attention to efficiently and explicitly model rich global features in the field of natural

* Corresponding author
E-mail addresses: zhuyu@ecust.edu.cn (Y. Zhu), yilei707@icloud.com (L. Yi).

https://doi.org/10.1016/j.cmpb.2023.107846
Received 27 June 2023; Received in revised form 3 October 2023; Accepted 4 October 2023
Available online 5 October 2023

language processing (NLP). Inspired by that, several Transformer-based networks have been proposed to improve the performance of medical image segmentation. Swin-Unet [12] modifies the Swin Transformer block into an UNet. Furthermore, TransFuse [13], TransUNet [1] and MedT [14] make attempts to fuse Transformers and CNNs for enhancing the ability of medical segmentation. The Transformers encoder is employed to capture the long-range dependency, while the CNN encoder can focus on local spatial contextual information.

Despite the advantages gained from combining CNN and Transformer, there are still a few limitations that need to be addressed. First, the heavy computational burden of the self-attention mechanism limits their practical application [15]. More recently, MLP-based architectures have attained competitive results with CNN and Transformer architectures. For instance, AS-MLP [16] is the first MLP-based network to be employed in image segmentation, capturing long-range dependencies in the image with a series of MLP blocks. We demonstrate three different architectures and compare their parameters and computational complexity in Fig. 1. Note that computational complexity is measured by floating-point operations per second (FLOPs). Here, H and W represent the length and width of the feature maps, respectively. And, P stands for patch size. K is kernel size, and C_i stands for channel size of input and output features, where $i \in \{1, 2, 3\}$. The computational complexity of self-attention is proportional to the square of the number of patches $\frac{HW}{P^2}$, while the computational complexity of MLP is proportional to the square of the channel size C_i. Generally, the patch number is much larger than the channel size. The MLP-based method used a series of convenient MLP blocks to replace the self-attention mechanism which can reduce the heavy computational burden. Therefore, a novel MLP-CNN based dual-path complementary network (MC-DC) is proposed for medical image segmentation. To light the computational burden, we introduce the multi-layer perceptron (MLP)-based methods to replace the Transformer-based method, which can achieve competitive results compared to Transformer, without using a self-attention mechanism. Specially, we design a dual-path complementary (DPC) module to effectively fuse multi-level features from MLP and CNN. Functionally, the MLP-based path can capture long-range information, while the CNN-based path helps to provide refined features for the corresponding features from MLP.

Second, most prior works solely focus on enhancing the performance of the feature encoder, with little emphasis placed on the design of the feature decoder. To integrate low-level features from the DPC module with high-level features learned by the encoder, we propose two feature fusion modules in the decoder stage. In detail, the designed cross-scale global feature fusion (CS-GF) module aims to rebuild global semantic information with the help of cross-scale attention. Meanwhile, the proposed cross-scale local feature fusion (CS-LF) module pays attention to reconstructing local spatial contextual information. Finally, we leverage a simple and efficient segmentation mask feature fusion (SMFF) module to combine the segmentation results of the dual-path decoder. The source codes have been uploaded at https://github.com/xiaobaimo/MC-DC for evaluation.

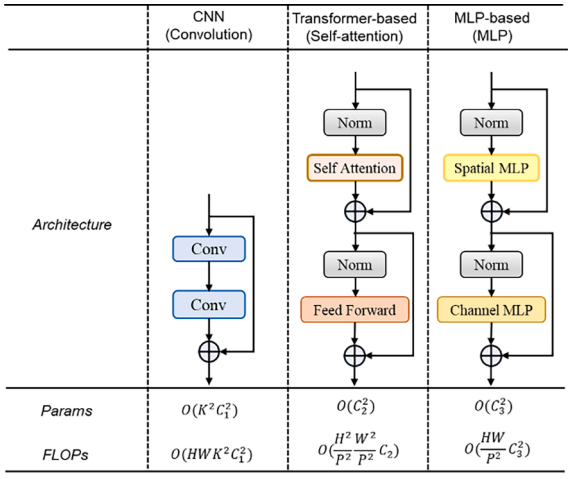

Fig. 1. Comparison of different architectures, along with their respective parameters and computational complexities.

To summarize, our main contributions can be outlined as follows:

(1) We propose a novel MLP-CNN based dual-path complementary network for medical image segmentation, MC-DC. The CNN encoder can focus on local spatial contextual information. The MLP encoder is employed to capture the long-range dependency, without using a complex self-attention mechanism.
(2) We design a dual-path complementary (DPC) module to effectively fuse multi-level features from MLP and CNN, which can efficiently aggregate the complementary information.
(3) The dual-path decoder is utilized to reconstruct global and local information, respectively. The cross-scale global feature fusion (CS-GF) module aims to rebuild global semantic information with the help of cross-scale attention, while the proposed cross-scale local feature fusion (CS-LF) module pays attention to reconstructing local spatial contextual information. In addition, we leverage a simple and efficient segmentation mask feature fusion (SMFF) module to combine the segmentation results of the dual-path decoder.
(4) Copious experiments were performed on three typical medical image segmentation tasks. The results demonstrate that the proposed MC-DC network has exceptional generalization capabilities, achieving higher results than other state-of-the-art methods while also exhibiting lower computational complexity.

2. Related work

2.1. U-Net and its variants for medical image segmentation

U-Net [2] is a pioneering work in medical image segmentation, utilizing a U-shaped architecture consisting of an encoder (downsampling path) and a decoder (upsampling path). The encoder is responsible for extracting high-level features from the medical image, while the decoder is used to map these features back to the original image size to generate the segmentation result. Specially, skip connections are utilized to connect each layer in the decoder to the corresponding layer in the encoder. This structure helps to preserve the details of the image and improves the accuracy of the segmentation. Inspired by that, many variants of U-Net are proposed to improve the performance of medical segmentation. For instance, UNet++ [3] is an extension of U-Net that incorporates a nested and dense skip pathway structure. It introduces a "deep supervision" mechanism, where the features generated by each skip layer are used to predict the segmentation mask at different scales. UNet3+ [4] is a further extension of Unet++, which introduces a multi-scale feature fusion module that fuses features from different scales to generate more comprehensive and diverse features. O-Net [17] is also based on U-Net, which is an architecture consisting of two convolutional autoencoders. Additionally, various attention-guided mechanisms have been designed to enhance the accuracy of segmenting objects of interest in medical images. Channel-Unet [7] incorporated spatial channel-wise convolution into the up-sampling and down-sampling modules, allowing for the extraction of mapping relationships of spatial information between pixels. Rca-u-net [8] integrated the U-Net architecture with residual channel attention blocks, thereby enhancing the network's capability to prioritize informative features and yield superior quantification results. [9] adopted squeeze-and-excitation block [18] after concatenation of low-level and high-level features to effectively enhance channel attention. [19] proposed a two-branch structure-guided segmentation network for 3D neuron reconstruction. Additionally, BAANet [20] incorporated both

channel attention and spatial attention into the U-Net architecture.

Although U-Net and its variants are capable of achieving remarkable segmentation results in medical images, they are still constrained by their limited receptive field of convolution kernels, which limits their capacity to capture wider global contextual information.

2.2. Transformer and CNN combined segmentation methods

To leverage the strengths of both CNN and Transformer, TransUNet [1] employed CNNs to extract the initial features of medical images, and subsequently utilized Transformers to further process the extracted features. After that, the decoder upsampled the encoded features and merged them with the high-resolution CNN feature maps to achieve precise localization. TransFuse [13] is a parallel approach that combines Transformers and CNNs, enabling the efficient capture of both global dependencies and low-level spatial details. Specially, the BiFusion module is proposed to efficiently merge the multi-level features from both branches. [14] proposed a novel architecture called the gated axial-attention model (MedT), which adds an additional control mechanism to the self-attention in existing architectures. Yuan et al. [21] designed a CTC–Net for medical image segmentation which combines ResNet34 [22] and Swin Transformer block [23]. Furthermore, the feature complementary module (FCM) is used for cross-wisely fusing features by a cross-domain fusion manner. These complementary models demonstrate significant advancements in medical segmentation compared to the pure CNN-based method. However, there are still a few limitations. First, self-attention mechanism presents a heavy computational burden that limits its practical application. Second, most prior works solely focus on designing the feature encoder, with little emphasis placed on the design of the feature decoder.

2.3. MLP-based architectures

More recently, MLP-based architectures have attained competitive results with CNN and Transformer architectures. Tolstikhin et al. [24] first presented the MLP-Mixer, which contains two types of MLP operation: mixing the per-location features and spatial information, respectively. Following this work, other MLP-based architectures such as

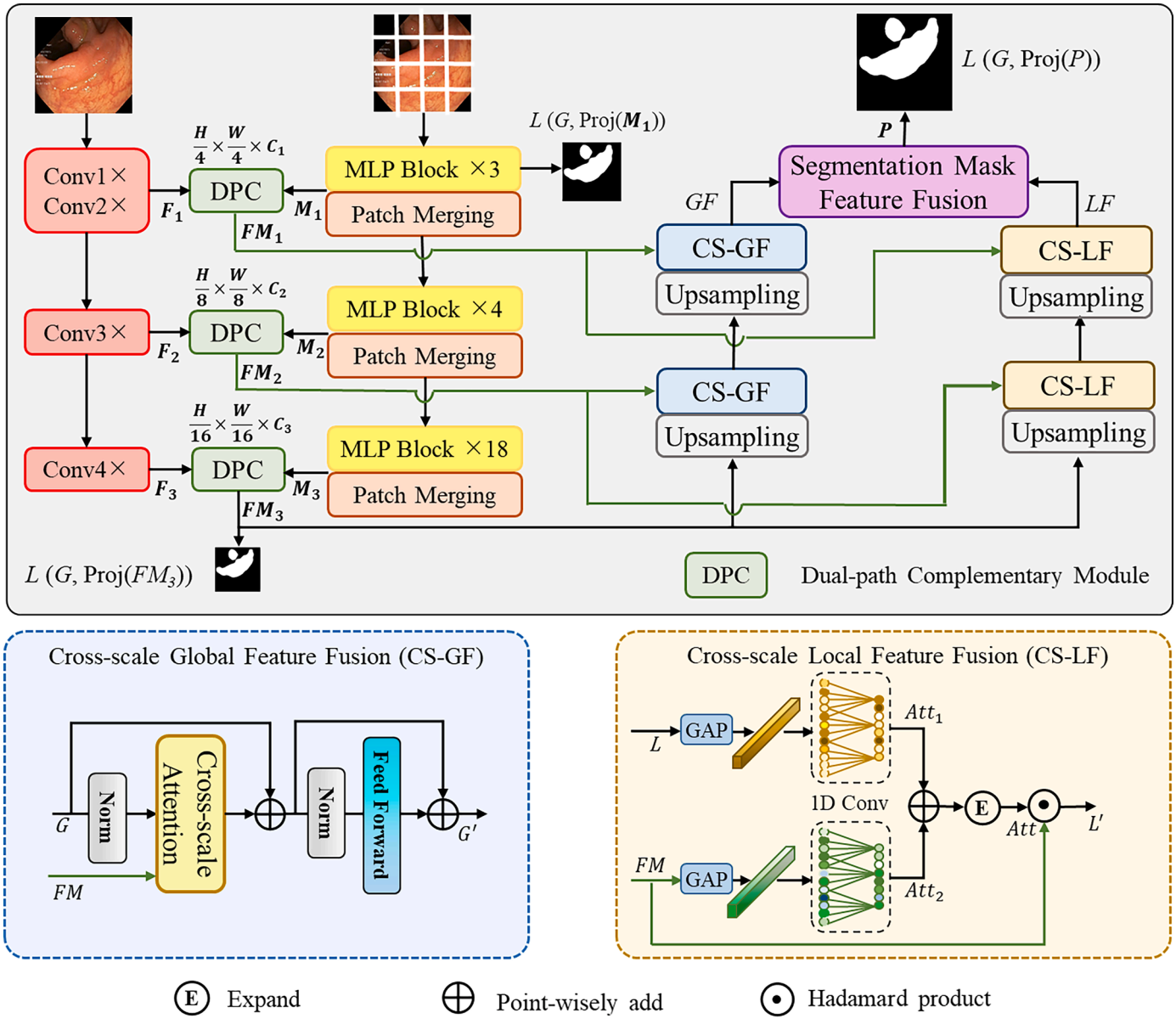

Fig. 2. Overview of the proposed MC-DC network, which is comprised of a dual-path encoder and a dual-path decoder. To achieve this, we employed Res2Net-50 and Wave-MLP as our dual-path encoder. In the decoder stage, the CS-GF module and CS-LF module are respectively designed to restore global and local information.

ResMLP [25] and gMLP [26] have been developed to further enhance performance. However, these architectures are dependent on image size and therefore may not be practical for tasks such as object detection and segmentation. To address this, AS-MLP [16] and CycleMLP [27] are proposed for dense prediction tasks, such as instance segmentation, and object detection. AS-MLP [16] designed an axial shift module to capture the information flow from different axial directions. CycleMLP [27] proposed a Cycle Fully-Connected Layer, which can deal with various image scales. In addition, Wave-MLP [28] transforms each token into a wave that possesses amplitude and phase, allowing for dynamic modulation of the relationship between tokens and the fixed weights in MLP.

To simultaneously leverage the advantages of CNN and MLP, we propose a novel dual-path complementary network. Furthermore, the DPC module is designed to effectively fuse multi-level features from MLP and CNN. In the decoder stage, we propose CS-GF and CS-LF to rebuild global semantic information and local spatial contextual information, respectively. Finally, we leverage a simple and efficient SMFF to combine the segmentation results of dual-path decoder.

3. Method

3.1. Architecture overview

The proposed MC-DC network for medical image segmentation mainly consists of a dual-path encoder and a dual-path decoder, as shown in Fig. 2. Given the input image $I \in \mathbb{R}^{H\times W\times C}$, we employed a CNN-based model (Res2Net-50 [29]) to capture three pyramidal features $F_i \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$, where $i \in \{1, 2, 3\}$ and $C_i \in \{64, 128, 320\}$. Meanwhile, an MLP-based model (Wave-MLP [28]) is utilized to extract M_i with the same scale. Then, the dual-path complementary (DPC) module is performed to effectively fuse the same scale features from CNN and MLP, yielding the fused features FM_i. After that, two trunk decoders gradually rebuild the high-level features FM_3 to original resolution with the help of CS-GF module and CS-LF module. Finally, we leverage a simple and efficient SMFF module to combine the segmentation results of two decoders.

3.2. Dual-path complementary module

To effectively fuse multi-level features from CNN and MLP, we proposed a Dual-path complementary (DPC) module, as shown in Fig. 3. First, global average pooling (GAP) followed by a 1D convolution with a kernel size of 5 is used as efficient channel-wished attention, computing as follows:

$$ATT_F = 1D - Conv(GAP(F)) \tag{1}$$

Then, the produced attention vector $ATT_F \in \mathbb{R}^{1\times 1\times c}$ and $ATT_M \in \mathbb{R}^{1\times 1\times c}$ are divided into n groups with length l, stood for $G_F \in \mathbb{R}^{n\times l}$ and $G_M \in \mathbb{R}^{n\times l}$, where k is equal to 16. After that, a fused matrix $R \in \mathbb{R}^{n\times n}$ is obtained through Eq. (2).

$$R = G_F G_M^T \tag{2}$$

The modulation factor $S_F \in \mathbb{R}^c$ and $S_M \in \mathbb{R}^c$ can formally be attained as follows:

$$S = Sigmoid(ATT + Linear(flatten(R))) \tag{3}$$

In addition, we utilized local attention (LA) to further enhance the expressiveness of local spatial contextual information of input feature maps F and M, gaining L_F and L_M. The core of LA is to use two 1×1 convolution operations to interact with features among different channels, as shown in Eq. (4).

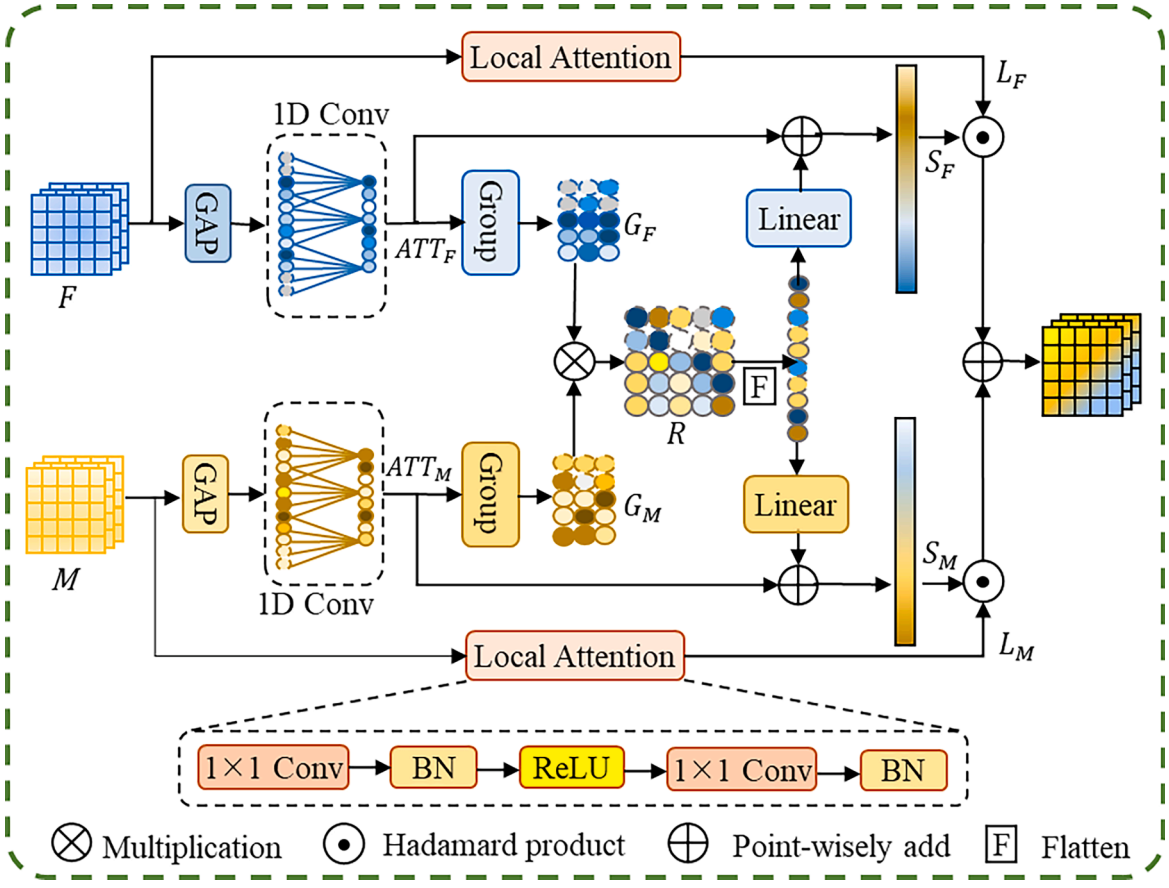

Fig. 3. Illustration of the dual-path complementary (DPC) module. Blue and yellow feature maps are yielded by Res2Net-50 and Wave-MLP, respectively.

$$L_F = BN(Conv(ReLu(BN(Conv(f)))))) \tag{4}$$

Finally, the fused feature maps *FM* is expressed as follows:

$$FM = S_M \cdot L_M + S_F \cdot L_F \tag{5}$$

3.3. Dual-path decoder

To respectively reconstruct global and local information, we designed a dual-path decoder that mainly consists of the cross-scale local feature fusion (CS-LF) module and the cross-scale global feature fusion (CS-GF) module. The details of the two modules are depicted in the blue and yellow blocks, as shown in Fig. 2.

CS-LF module pays attention to reconstructing local spatial contextual information. We first upsample the output features of the encoder via bilinear interpolation and obtain upsampled features $L \in \mathbb{R}^{h \times w \times c}$. Then, global average pooling (GAP) followed by a 1D convolution with a kernel size of 5 is used as efficient channel-wished attention and attains attention vector $Att_1 \in \mathbb{R}^{1 \times 1 \times c}$, which can be described as Eq. (6).

$$Att_1 = 1D - Conv(GAP(L)) \tag{6}$$

In addition, we introduce low-level features $FM \in \mathbb{R}^{h \times w \times c}$ to complement boundaries and spatial structure. Att_2 is obtained through the same operation. After that, we add $Att_1$with Att_2 and expand to the original shape, yielding $Att \in \mathbb{R}^{h \times w \times c}$. Finally, the output of the CS-GF module can be produced by Eq. (7).

$$L^{'} = Att \cdot FM \tag{7}$$

CS-GF module is utilized to rebuild global semantic information. The kernel component of the CS-GF module is cross-scale attention (CSA), which is described in Fig. 4 in detail. The inputs of CSA are from upsampled feature $G \in \mathbb{R}^{h \times w \times c}$ and low-level feature $FM \in \mathbb{R}^{h \times w \times c}$. First, feature *G* is reshaped to 2D tokens and passes through three linear projections, yielding $Q \in \mathbb{R}^{hw \times c}$, $K \in \mathbb{R}^{hw \times c}$, and $V \in \mathbb{R}^{hw \times c}$. Meanwhile, the *FM* performs the same operation and attains $D \in \mathbb{R}^{hw \times c}$. After conducting two similarity comparisons, CSA is capable of establishing a strong and global correlation between high-level features and low-level features. The following equations can describe the calculation process.

$$QD = Softmax\left(\frac{QD^T}{\sqrt{d}}\right) \tag{8}$$

$$DK = Softmax\left(\frac{DK^T}{\sqrt{d}}\right) \tag{9}$$

$$Y = QD \times (DK \times V) \tag{10}$$

3.4. Segmentation mask feature fusion module

To effectively fuse the segmentation results from the dual-path decoder, we propose a segmentation mask feature fusion (SMFF) module, the detailed structure is illustrated in Fig. 5. The inputs of the module are local feature maps $LF \in \mathbb{R}^{h \times w \times c}$ yielded by CS-LF module and global feature maps $GF \in \mathbb{R}^{h \times w \times c}$ produced by CS-GF module. We first directly add *L* with *G*, attaining mixed features $GL \in \mathbb{R}^{h \times w \times c}$. Then, the local attention and global attention are performed on the mixed features *GL* respectively, and then enhanced feature maps are added as a weighted attention map $Att \in \mathbb{R}^{h \times w \times c}$. This process can be expressed by Eq. (11).

$$Att = Loc(GL) + Glob(GL) \tag{11}$$

Among them, $Loc(\cdot)$ and $Glob(\cdot)$ represent local attention operation and global attention operation respectively, which can be expressed as Eq. (12) and Eq. (13):

$$Loc(GL) = BN(Conv(ReLU(BN(Conv(GL))))) \tag{12}$$

$$Glob(GL) = BN(Conv(ReLU(BN(Conv(GAP(GL))))))) \tag{13}$$

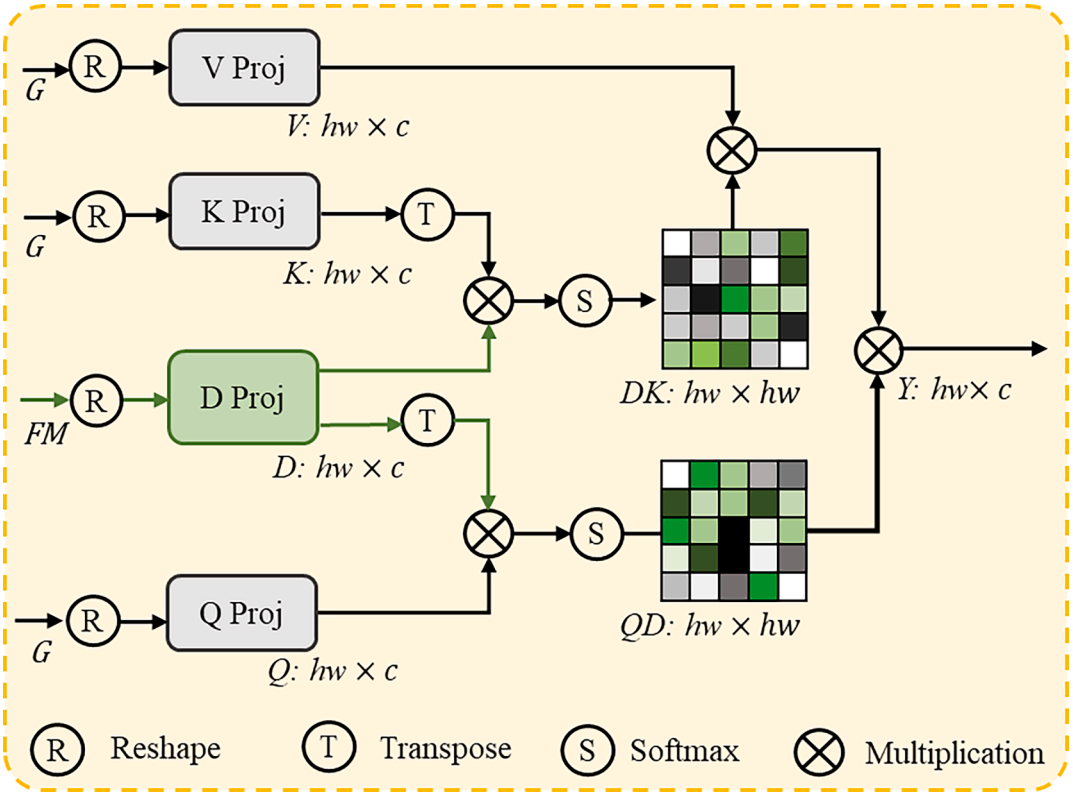

Fig. 4. The architecture of cross-scale attention (CSA).

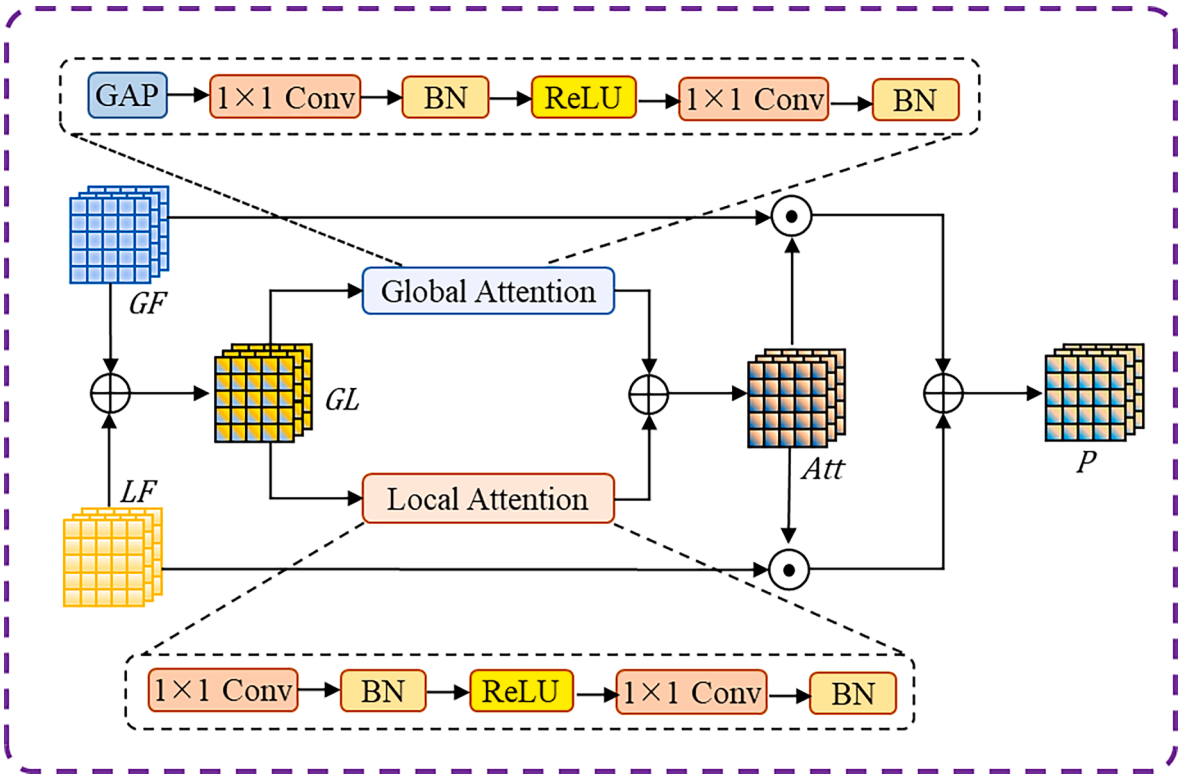

Fig. 5. Segmentation mask feature fusion (SMFF) module.

where *Conv* is a convolution operation of 1 × 1, *ReLU* is an activation function, *BN* refers to a batch normalization operation, and *GAP* refers to a global average pooling. The core of local attention is to use two 1 × 1 convolution operations to interact with features between different channels and enhance the expressiveness of local features. Global attention aims to extract global information from the network and enhance the expression ability of global features. Then, the weighted attention map *Att* is respectively multiplied by the global feature maps *G* and local feature maps *L* to increase the weight of important regions in the feature map. Finally, the enhanced dual-path features are added together as the output of the SMFF module. This process can be expressed by Eq. (14).

$$P = Att \cdot GF + Att \cdot LF \tag{14}$$

3.5. Loss function

The entire network uses a combination of intersection over union (IoU) loss and binary cross-entropy (BCE) loss for end-to-end training that can be described as follows:

$$\mathscr{L} = \ell_{IoU} + \ell_{bce} \tag{15}$$

Where, ℓ_{IoU} is IoU loss, ℓ_{bce} stands for BCE loss. The mixed loss employed in the training process can restrict the prediction map at both the object level and pixel level. Specially, deep supervision [30] is leveraged to enhance the gradient flow, particularly by providing additional supervision to the last fusion feature maps (FM_3). Therefore, the final training loss is given in Eq. (16),

$$\mathscr{L} = \alpha\mathscr{L}(G, Proj(M_1)) + \beta\mathscr{L}(G, Proj(FM_3)) + \gamma\mathscr{L}(G, Proj(P)) \tag{16}$$

where $\alpha = 0.2$, β=0.3, and $\gamma = 0.7$ are weighted hyperparameters and G is groundtruth. Note that *Proj* is a segmentation head that can restore the feature maps to the groundtruth size.

4. Experiments

To assess the learning and generalization abilities of our MC-DC network, we performed experiments on three typical medical image segmentation tasks: skin lesions segmentation, polyp segmentation, and lung lesions segmentation. We first briefly provide a concise overview of all the datasets. Then, the implementation and evaluation are presented. In the end, we compare our results with other state-of-the-art (SOTA) methods that have been recently published and provide a comprehensive analysis of the different components of our proposed method through detailed ablation studies.

4.1. Datasets

Skin lesions segmentation. We adopt ISIC 2018 dataset [31] for this task, which consists entirely of microscope RGB images, and the segmentation labels are manually labeled by professional clinical doctors on the area of skin disease lesions. This dataset has already been officially divided into 2,594 for training and 100 for testing. In addition, the PH2 dataset [32], which consists of 200 images with skin lesions, is used as a supplementary test to verify the generalization of the model. Note that all images are consistently adjusted to 192 × 256.

Polyp segmentation. We adopt two public polyp datasets including Kvasir-SEG [33] and CVC-ClinicDB [34] for this task. The Kvasir-SEG and CVC-ClinicDB dataset has been widely used in research on the detection and classification of gastrointestinal diseases, contributing to the development of new algorithms and techniques. Following [35], the Kvasir-SEG dataset is randomly assigned 880 images for training and 120 for testing, while the CVC-ClinicDB dataset comprises 550 images for training and 62 for testing. Each image is resized into 256 × 256.

Lung lesions segmentation. The COVID-DS36 dataset, jointly established by our collaborating hospitals, is utilized in this task. The dataset consists of 4369 computed tomography (CT) images obtained from lung scans of 36 patients, of which 18 patients were diagnosed with COVID-19 infection and the remaining 18 were healthy individuals. As stated in

[36,37], the clinical diagnosis proves that lung CT images exhibit evident imaging characteristics of COVID-19. 18 patients presented varying symptoms, and their CT images were annotated by medical professionals, revealing the presence of three diseases: ground-glass opacity (GGO), lung consolidation, and interstitial infiltration. We separated the data according to patients to avoid extreme similarity between the training and testing sets. In detail, the dataset is randomly assigned 28 patients (3,496 images) for training (80%), and 8 patients (873 images) for testing (20%), with an image size of 224 × 224.

4.2. Implementation and evaluation

Implementation details. We implemented our MC-DC network using Python 3.8 and PyTorch 1.7. All experiments were conducted on a PC equipped with an Intel(R) Core(TM) i9-10940X CPU and an Nvidia GTX 3090 with 24GB of memory. Adam optimizer with a learning rate of 1e-5 was utilized to update the parameters of networks, and the cosine schedule [38] was used for weight decay. The network was trained for 100 epochs with a batch size of 8. Additionally, we employed random horizontal flips, random vertical flips, and random scale rotation shifts as data augmentation methods. Res2Net-50 [29] is adopted as the CNN-based encoder, while Wave-MLP [28] is utilized as the MLP-based encoder. Note that both models are initialized with pre-training weights from ImageNet-1k.

Evaluation metrics. To evaluate the performance of our MC-DC network, we utilize two classification metrics (recall and precision) and four segmentation metrics which include average dice coefficient (Dice), average Intersection over Union (IoU), average Hausdorff distance (HD), and average symmetric surface distance (ASSD). The six metrics can be expressed as follows:

$$Rec = \frac{TP}{TP + FN} \tag{17}$$

$$Pre = \frac{TP}{TP + FP} \tag{18}$$

$$Dice = \frac{2(P \cap G)}{|P| + |G|} \tag{19}$$

$$\mathrm{IoU} = \frac{P \cap G}{P \cup G} \tag{20}$$

$$\mathrm{HD} = \max\left(\underset{p\in P}{max}\{\underset{g\in G}{min}\| p-g \|\}, \underset{g\in G}{max}\left\{\underset{p\in P}{min}\| g-p \|\right\}\right) \tag{21}$$

$$ASSD = \frac{1}{|P| + |G|}\left(\sum_{p\in P}\underset{g\in G}{min}||p-g|| + \sum_{g\in G}\underset{p\in P}{min}||g-p||\right) \tag{22}$$

where *TP, FN* and *FP* stand for true positives, false negatives and false positives, respectively. *P* and *G* represent the predicted area and ground truth, respectively. Recall (Rec) measures the probability of not missing true positive cases, while precision (Pre) measures the probability of diagnosing true positive cases. *Dice* and *IoU* both measure of spatial overlap between the predicted masks and the ground truth. The higher score, the closer the prediction result is to the ground truth. Meanwhile, HD and ASSD are utilized for measuring the boundary surface distance between the predicted results and the ground truth. Note that predicted results and ground truth are closer with a lower score.

4.3. Comparison with state-of-the-art methods

4.3.1. Experiments on skin lesions segmentation

To evaluate the skin lesions segmentation performance of our MC-DC, 16 recent SOAT models are used for comparison, including 7 pure CNN-based methods and 9 CNN-Transformer based methods. Table 1 lists the results of comparative models and our MC-DC network on ISIC2018 measured by the Iou, Dice, ASSD, and HD. It can be observed that the CNN-based methods are unable to overcome the bottleneck caused by their limited capability to capture long-range dependencies. Meanwhile, most CNN-Transformer based methods outperform CNN-based methods, demonstrating the positive role of combining the CNN and Transformer. Compared with the original U-Net [2], BAT [39] can improve the IoU and Dice by 6.51% and 3.70%, respectively. Moreover, the proposed MC-DC network can achieve the SOTA results in all listed metrics (85.32% IoU, 91.69% Dice, 9.52mm ASSD and 24.73mm HD), which highlights the beneficial impact of integrating the MLP and CNN techniques and dual-path decoder.

Table 1
Comparison with different SOTA models on ISIC2018. For each column, the best results are highlighted in bold.

Model	IoU (%)↑	Dice (%)↑	ASSD (mm)↓	HD (mm)↓
U-Net [2]	77.86	87.55	17.58	41.28
U-Net++ [3]	78.31	87.83	17.27	42.75
Attention U-Net [40]	78.43	87.91	16.79	41.90
ResUNet [41]	78.60	86.20	-	-
DeepLabV3+ [42]	80.62	88.49	15.35	34.74
CE-Net [43]	81.65	89.17	14.76	31.01
CA-Net [44]	82.73	89.31	14.82	32.47
Swin U-Net [12]	83.46	90.78	10.88	28.60
MedT [14]	81.78	87.92	15.27	32.40
TransFuse [13]	80.63	89.27	12.40	28.37
FAT-Net [45]	82.02	89.03	-	-
TransUNet [1]	82.20	89.40	13.17	32.05
TransNorm [46]	84.40	90.87	11.46	27.08
CASF-Net [35]	84.1	90.8	-	-
SwinPA-Net [47]	85.4	91.1	-	-
BAT [39]	84.37	91.25	9.92	27.69
MC-DC (ours)	**85.32**	**91.69**	**9.52**	**24.73**

Fig. 6 illustrates the visual comparisons between the prediction results acquired from our MC-DC network and the compared methods. To enhance the visibility of the segmentation results in contrast to the lesion region edges, we transformed the segmentation results into translucent masks and fused them with the original image for better visualization. The first row demonstrates that our model can accurately detect lesions covered by the hair with the highest accuracy. The first and second rows prove that our MC-DC network possesses exceptional capabilities in capturing small-scale lesions. From the last two lines, it can be observed that our MC-DC network also can accurately detect the large-scale lesions. In summary, the proposed MC-DC network has superior performance than other compared methods in handling complex cases with varying scales and blurred boundaries.

To further verify the generalization performance of our MC-DC network, we utilized the PH2 dataset as a supplementary test set, which was not included in the training process. Fig . 7 illustrates the results between HD and Dice of our MC-DC network and the compared methods. Based on Fig. 7, our method achieves the highest Dice score and the lowest HD score, indicating that it produces predicted results that are closest to the ground truth for both the interior and edge of the lesion. Therefore, it can be demonstrated that our MC-DC network has good generalization and robustness.

4.3.2. Experiments on polyp segmentation

Table 2 lists the results of comparative models and our MC-DC network on Kvasir-SEG and CVC-ClinicDB datasets measured by the Dice, Iou, Rec, and Pre. As can be observed from the table, our MC-DC network outperforms the latest both CNN methods (i.e., HarDNet-MSEG [48] and EU-Net [49]) and CNN-Transformer based methods (i.e., DS-TransUnet [50] and CASF-Net [35]) in almost all metrics. In detail, our MC-DC network attains 91.6% Dice, 93.8% Rec, and 91.4 % Pre, improving by 0.2%, 0.2%, and 0.1% over the second place method (CASF-Net) on the Kvasir-SEG dataset. Again, the proposed MC-DC network achieves the best results (94.4% Dice) on CVC-ClinicDB

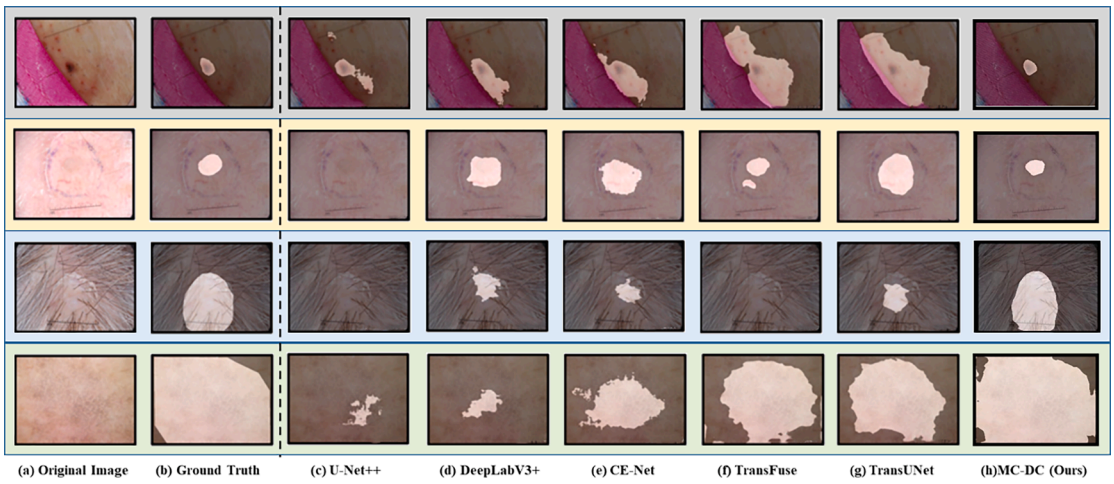

Fig. 6. Comparison of visual skin lesions segmentation results. The first and second column stand for the original image and ground truth, respectively. (c)-(h) recovered results from U-Net++, DeepLabV3+, CE-Net, TransFuse, TransUNet, and our MC-DC, respectively.

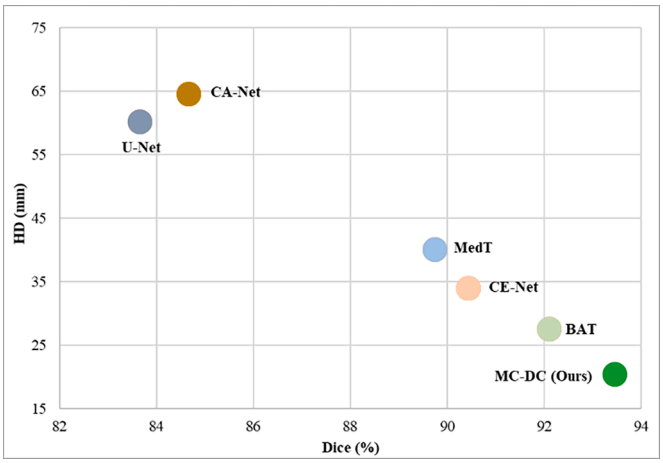

Fig. 7. Generalizability results using PH2 dataset. The X-axis represents the intervals of Dice, and the Y-axis stands for the intervals of HD. Our MC-DC network is represented by green circles.

datasets.

Furthermore, we illustrate the visual comparisons of polyp segmentation between the prediction results obtained from our MC-DC network and the compared methods on Fig. 8. The first and the second column represent the original polyp images and the ground truth, respectively. The qualitative segmentation results of the proposed MC-DC network and the comparison networks are illustrated as follows. The predicted masks generated by our MC-DC network outperform other models as it closely resembles the boundary and shape of the ground truths.

4.3.3. Experiments on lung lesion segmentation

As shown in Table Table 3, we employed five CCN methods (U-Net [2], UNet++ [3], EU-Net [49], PSPNet [52] and SegNet [53]) and four CNN-Transformer based methods (TransFuse [13], TransUNet [1], Swin-Unet [12] and CASF-Net [35]) as the comparison networks to evaluate the performance on lung lesion segmentation. Here, the COVID-DS36 dataset is partitioned into five equal parts and the 5-fold cross-validation is employed to develop a more generalized model. Note that the quantitative data were presented as values [95% confidence interval]. The classical U-Net achieves 76.1% [69.2%, 83.1%] Dice, 82.3% [76.3%, 88.4%] Dice, and 80.64% [77.2%, 84.0%] Dice on GGO, interstitial infiltrates, and lung consolidation, respectively. In addition, we find that UNet++ can obviously improve the Dice metric of the three lesion types by 10.9%, 8.7% and 10.0%, respectively. However, the Rec metrics obtained by UNet++ show a slight decrease. The same issue has also been observed in EU-Net, PSPNet, and SegNet. On the contrary, TransUNet, Swin-Unet, and CASF-Net have improved both the Dice and Rec metrics. The Dices obtained by Swin-Unet are 86.9% [86.1%, 87.7%], 91.1% [90.6%, 91.6%] and 91.7% [91.3%, 92.1%] on the three lesion types, respectively. Meanwhile, the Recs are 94.5% [93.9%, 94.9%], 97.0% [96.6%, 97.4%], 96.4% [96.1%, 96.7%]. Based on Table 3, it is clear that the proposed MC-DC network outperforms all comparative networks in almost all metrics, with Dice scores of 87.6%, 92.3%, and 92.3%, and Recall scores of 96.4%, 97.6%, and 97.1%.

In addition, we also illustrate the average Dice of three lesion types using box plot on Fig. 9. From that, we can clearly find that our MC-DC can achieve the highest average dice, and the strongest model robustness due to its most concentrated data distribution. Specially, we still use paired t-test for statistical significance testing and report the P-values. In general, a resulting p-value below 0.05 is considered acceptable. It is

Table 2
Comparison with different SOTA models on polyp segmentation. For each column, the best results are highlighted in bold.

Method	Kvasir-SEG				CVC-ClinicDB			
	Dice (%)↑	IoU (%) ↑	Rec (%) ↑	Pre (%) ↑	Dice (%)↑	IoU (%) ↑	Rec (%) ↑	Pre (%) ↑
U-Net [2]	74.8	64.0	82.1	76.3	77.9	69.6	78.9	80.2
DoubleU-Net [51]	81.3	73.3	84.0	86.1	92.4	86.1	84.6	90.7
HarDNet-MSEG [48]	90.4	84.8	92.3	90.7	92.4	86.1	90.0	92.0
EU-Net [49]	90.8	85.4	93.4	90.9	90.2	84.6	90.6	87.8
TransUNet [1]	89.8	86.3	91.2	91.3	92.3	86.9	94.2	91.7
Swin-Unet [12]	89.0	82.5	90.6	90.6	90.6	84.9	91.8	90.7
DS-TransUnet [50]	91.3	85.9	93.6	91.6	94.2	89.4	95.0	93.7
CASF-Net [35]	91.4	**87.1**	93.6	91.3	93.4	89.9	95.2	**94.2**
MC-DC (ours)	**91.6**	86.7	**93.8**	**91.4**	**94.4**	**90.0**	**95.2**	94.1

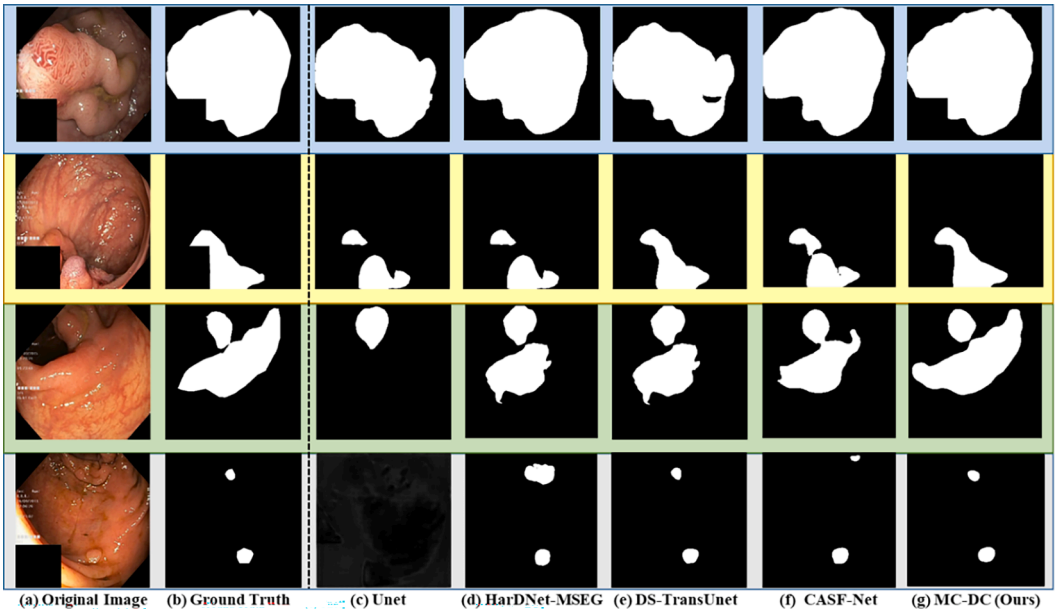

Fig. 8. Comparison of visual polyp segmentation results. The first and second column stand for the original image and ground truth, respectively. (c)-(g) recovered results from U-Net, HarDNet-MSEG, CASF-Net, and our MC-DC, respectively.

Table 3
Comparison with different SOTA models on lung lesion segmentation. For each column, the best results are highlighted in bold.

Method	GGO		Interstitial Infiltrates		Consolidation	
	Dice (%)↑	Rec (%)↑	Dice (%)↑	Rec (%)↑	Dice (%)↑	Rec (%)↑
U-Net [2]	76.1 [69.2, 83.1]	94.8 [94.1, 95.5]	82.3 [76.3, 88.4]	97.2 [96.8, 97.6]	80.6 [77.2, 84.0]	96.3 [96.0, 96.5]
UNet++ [3]	87.0 [86.3, 87.7]	93.9 [93.1, 94.6]	91.0 [90.3, 91.7]	95.9 [95.5, 96.4]	90.6 [88.5, 92.7]	95.6 [95.3, 95.9]
EU-Net [49]	83.4 [83.0, 83.8]	93.2 [92.7, 93.7]	90.5 [90.1, 90.9]	95.3 [94.8, 95.8]	90.2 [89.7, 90.7]	95.5 [95.0, 96.0]
PSPNet [52]	84.1 [83.6, 84.6]	90.6 [90.0, 91.2]	89.7 [89.2, 90.2]	95.0 [94.0, 95.9]	90.5 [89.9, 91.1]	94.0 [93.4, 94.4]
SegNet [53]	83.8 [82.5, 85.1]	93.6 [92.5, 94.7]	91.6 [91.1, 92.1]	96.2 [95.9, 96.5]	89.7 [89.3, 90.1]	96.0 [95.7, 96.3]
TransFuse [13]	83.3 [82.4, 84.2]	93.3 [92.5, 94.1]	88.0 [87.5, 88.5]	95.0 [94.2, 95.8]	89.4 [88.6, 90.2]	95.0 [94.3, 95.7]
TransUNet [1]	86.0 [86.1, 87.7]	95.0 [94.5, 95.5]	91.1 [90.7, 91.5]	96.7 [96.2, 97.2]	90.4 [90.0, 90.8]	96.6 [96.1, 97.1]
Swin-Unet [12]	86.9 [86.1, 87.7]	94.5 [93.9, 94.9]	91.1 [90.6, 91.6]	97.0 [96.6, 97.4]	91.7 [91.3, 92.1]	96.4 [96.1, 96.7]
CASF-Net [35]	85.5 [85.0, 86.0]	94.5 [94.1, 94.9]	89.9 [89.5, 90.3]	96.3 [95.9, 96.7]	91.1 [90.7, 91.4]	96.7 [96.2, 97.2]
MC-DC (ours)	**87.6 [87.1, 88.1]**	**96.4 [96.0, 96.7]**	**92.3 [92.1, 92.6]**	**97.6 [97.2, 98.0]**	**92.3 [91.8, 92.7]**	**97.1 [96.8, 97.4]**

evident that all outcomes are below 0.05 on COVID-DS36 dataset, thus validating the statistical significance of our results. Specially, the visual comparisons of lung lesion segmentation are shown in Fig. 10. We use white dashed circles to indicate the improved segmentation performance of the proposed MC-DC network. For instance, the last sample has

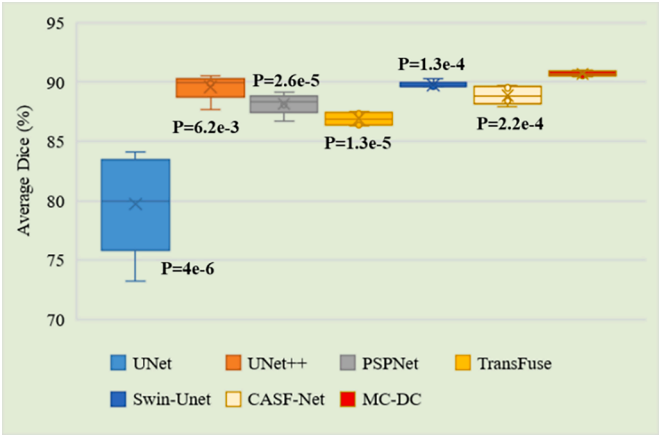

Fig. 9. Box plot of average Dice produced by different models. We also use paired t-test for statistical significance testing and illustrate the P-values.

three lesion types and the segmentation areas are relatively discrete. The prediction results yielded by the proposed MC-DC network are more closely aligned with the ground truths and exhibit superior segmentation capabilities in complex lesion areas when compared to other networks.

4.4. Ablation studies

In this section, we design comprehensive ablation studies to assess the effectiveness of each component in MC-DC network. The proposed MC-DC network is composed of dual-path encoder and decoder. Hence, we first design various combinations of encoder and decoder and conduct experiments on ISIC2018 dataset. As listed in Table 4, "CNN" is a CNN-based model (Res2Net-50), and "MLP" is an MLP-based model (Wave-MLP). "CS-GF" stands for global reconstruction decoder, while "CS-LF" represents local reconstruction decoder. From the first and second lines, it can be inferred that there is a significant performance gap when solely using CNN as the decoder. The Dice decreases by 1.14% dice and the ASSD increases by 2.45%. Then, it can be observed that MLP-based encoder can improve the results, benefitting from the ability

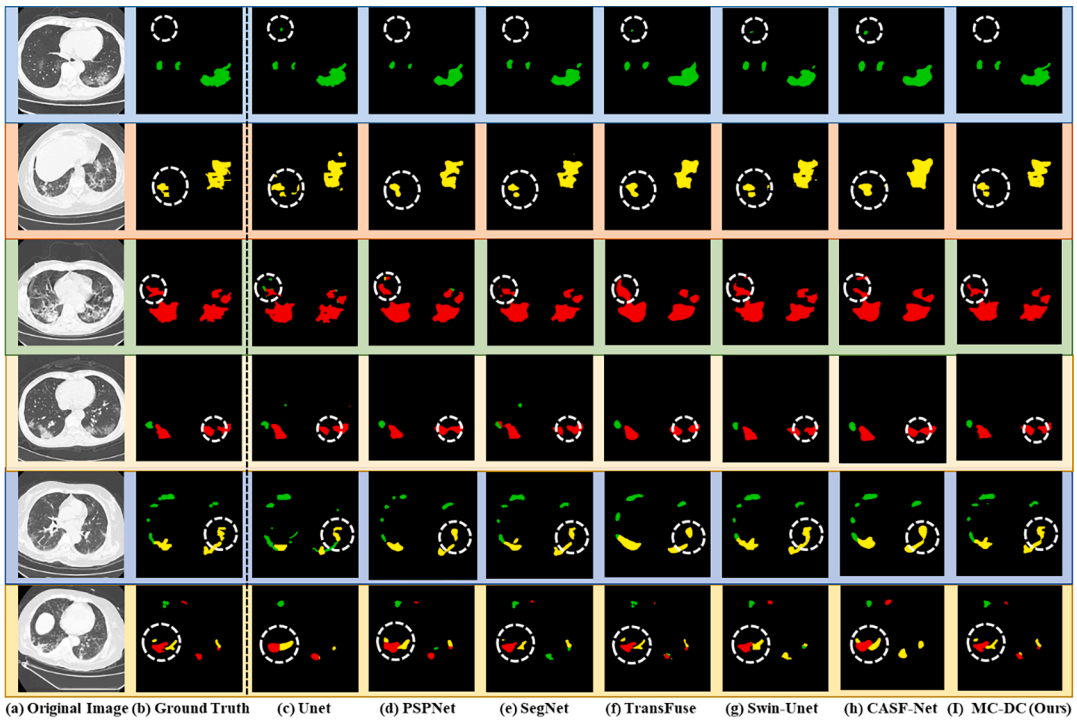

Fig. 10. Comparison of visual lung lesion segmentation results. The first and second column stand for the original image and ground truth, respectively. (c)-(I) recovered results from U-Net, PSPNet, SegNet, TransFuse, Swin-Unet, CASF-Net, and our MC-DC, respectively. The three color markers represent the three lesions, with green indicating ground-glass opacity (GGO), yellow indicating interstitial infiltration, and red indicating lung consolidation.

Table 4
Ablation studies on various combinations of encoder and decoder using ISIC2018 dataset.

Encoder		Decoder		Dice (%)↑	ASSD (mm) ↓
CNN	MLP	CS-GF	CS-LF		
√		√		90.55	11.97
√			√	90.29	12.38
	√	√		91.08	11.27
	√		√	90.83	11.57
√	√	√		91.38	10.46
√	√		√	91.29	10.73
√	√	√	√	91.69	9.52

to capture the global information. Moreover, MLP-CNN based dual-path encoder exhibits outstanding performance in medical image segmentation which can focus on both local spatial contextual information and long-range dependency. Furthermore, the decoder is also crucial. Our MC-DC network can surpass all variant networks when employing both local and global reconstruction decoder.

In addition, we also investigate the impact of the proposed dual-path complementary (DPC) module and segmentation mask feature fusion (SMFF) on COVID-DS36 dataset. The results are shown in Table 5. We first remove the SMFF module and directly add the two feature maps from two decoders. The average Dice decreases by 0.6% and the average Rec decreases by 0.8. Meanwhile, the Dice decreases by 0.9% without DPC. Specially, the average Dice and average Rec respectively decrease by 1.5% and 1.6% when removing both DPC and SMFF. It can be observed that our DPC and SMFF can efficiently aggregate the complementary information.

Table 5
Ablation studies on feature fusion modules using COVID-DS36 dataset. The "w/o" stands for "without".

Methods	Average Dice (%)↑	Average Rec (%)↑
w/o SMFF	90.1 [89.8, 90.4]	96.2 [95.8, 96.5]
w/o DPC	89.8 [89.5, 90.1]	95.9 [95.6, 96.2]
w/o (DPC+SMFF)	89.2 [88.8, 89.6]	95.4 [95.1, 95.7]
DPC+SMFF	90.7 [90.5, 90.9]	97.0 [96.9, 97.1]

4.5. Visualization of feature maps from dual-path encoder and decoder

In this section, we will comprehensively investigate the qualitative results of dual-path encoder and decoder. As shown in Fig. 11, we visualize feature maps generated by encoder (purple block) and decoder (green block) on three cases of COVID-DS36 dataset. Note that the lower response is demonstrated in blue while the higher are highlighted in red which can highlight areas of interest or concern. For the encoder, we first list the feature map F_3 and feature map M_3 generated by CNN branch and MLP branch, respectively. The details can be seen in Fig. 11 (b) and Fig. 11 (c). From feature map F_3, we can observe that the red areas are relatively small, indicating that the CNN is specifically attentive to local spatial contextual information. From the feature map M_3, it is evident that the red areas are significantly larger, implying that the MLP is capable of capturing long-range dependencies. However, it is worth noting that the majority of the areas of concern do not correspond to lung lesions. Furthermore, we also show the fused feature map FM_3 in Fig. 11 (d). We found that the red areas are more concentrated in lung lesions. It is evident that the proposed Dual-Path Complementary (DPC) module can accurately combine lung lesion information from the CNN and MLP branches.

For the decoder, we first list the feature map *LF* and feature map *GF* reconstructed by CS-LF module and CS-GF module, respectively. From Fig. 11 (f) and Fig. 11 (g), we can find that the proposed CS-LF module aims to reconstruct local spatial contextual information, while the designed CS-GF focuses on rebuilding global semantic information. Specially, we further visualize the feature map *P* that combines the output of the two branches which are more similar to the ground truths. The results prove the effectiveness of the SMFF module to combine the segmentation results of the dual-path decoder.

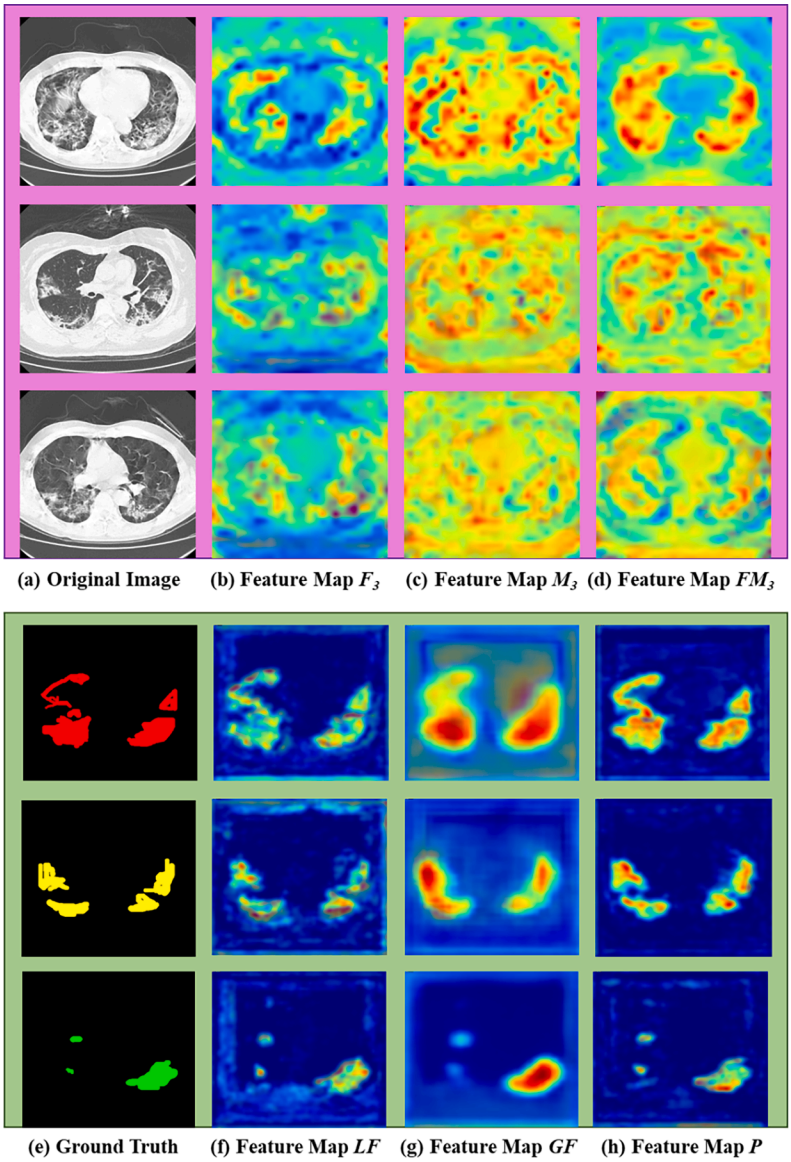

Fig. 11. Visualization of feature maps generated by an encoder (purple block) and decoder (green block) on three cases of COVID-DS36 dataset.

4.6. Comparison of inference efficiency

To assess the inference efficiency, we calculate the parameters and computational complexity of our MC-DC network and its variants on ISIC 2018 dataset. Here, we replace the MLP branch with the Transformer branch (PVTv2-B2 [54] and PVTv2-B3 [54]) in the encoder which we name as "Variant 1" and "Variant 2", respectively. In addition, we also list the inference efficiency of three SOTA methods: ResUNet [41], TransUNet [1], and BAT [39]. The results are shown in Table 6. Note that floating-point operations per second (FLOPs) are utilized to measure computational complexity. Based on Table 6, we can obtain the following observations: 1) ResUNet [41] has the highest Flops but the lowest Dice score. 2) TransUNet [1] has the highest number of parameters but a lower Dice score compared to our MC-DC. Meanwhile, BAT [39] has a lower number of parameters and Dice scores than our MC-DC. 3) Our MC-DC network has lower Flops compared with the “Variant 1” which has almost the same parameters. In addition, our MC-DC network can improve by 0.31% Dice over "Variant 2" which has higher parameters.

It proves that the MLP-based method used a series of convenient MLP blocks to replace the self-attention mechanism which can reduce the heavy computational burden and improve performance compared to other SOTA methods.

Table 6
Comparison of model efficiency on ISIC dataset. Note that the inputs are set to 192×256.

Method	Params (M)	Flops (G)	Dice (%)
ResUNet [41]	62.74	94.56	86.20
TransUNet [1]	105.32	38.52	88.91
BAT [39]	46.23	44.98	91.20
Variant 1	62.60	42.15	91.12
Variant 2	78.48	45.21	91.38
MC-DC (Ours)	65.93	41.13	91.69

5. Discussion

Recently, Transformers have leveraged self-attention to efficiently and explicitly model rich global features in the field of natural language processing (NLP). Hence, How to adapt Transformer architecture into medical segmentation has garnered increasing attention. Among that, the fusion of the CNN and Transformer in the encoder has achieved

remarkable performance. However, there are obvious limitations to such models that require addressing. First, the utilization of Transformer leads to heavy parameters, and its intricate structure demands ample data and resources for training. Second, most previous research had predominantly focused on improving the performance of the feature encoder, with little emphasis placed on the design of the feature decoder. To address these issues, we propose a novel MLP-CNN based dual-path complementary (MC-DC) network for medical image segmentation. After conducting sufficient experiments, we make detailed discussions as follows:

(1) As can be observed from Fig. 11, we can find that MLP architecture can also capture the global features in medical images. Furthermore, MLP architecture is simpler and more efficient rather than complex Transformer architecture. The quantitative comparison is also listed in Table 6.
(2) Extensive experiments were conducted on three typical medical image segmentation tasks. For skin lesions segmentation, our MC-DC network has superior performance than other compared methods in handling complex cases with different scales and blurred boundaries. From Table 2 and Fig. 8, we can find that the proposed MC-DC network has achieved the highest Dice 91.6% on Kvasir-SEG, and 94.4% Dice on CVC-ClinicDB datasets. Lung lesion segmentation is a multi-class segmentation task. Based on Fig. 10, the prediction results yielded by the proposed MC-DC network are more closely aligned with the ground truths and exhibit superior segmentation capabilities in complex lesion areas when compared to other networks.
(3) We design comprehensive ablation studies to assess the effectiveness of each component in MC-DC network. Based on Table 4, we find that the dual-path encoder and the dual-path decoder can achieve the best result. Table 5 proves the effectiveness of the SMFF module and DPC module. Furthermore, the visualization of feature maps generated by encoder and decoder more intuitively illustrates the reasonableness of the designed components.

Despite the promising clinical prospects of our MC-DC network's segmentation advantages, there are still some remaining shortcomings. First, the parameters of the networks needed to be further reduced. Second, the model also requires a large amount of data to enhance robustness. In the future, we will continue to investigate and develop more lightweight and efficient networks for medical image segmentation.

6. Conclusion

In this study, a novel MLP-CNN based dual-path complementary (MC-DC) network is proposed for medical image segmentation. It replaces the complex Transformer with cost-effective MLP. And, the dual-path decoder is designed to respectively reconstruct global and local information with the help of CS-GF module and CS-LF module. Specifically, the DPC module is proposed to effectively fuse multi-level features from MLP and CNN, and the SMFF module is leveraged to merge the segmentation outcomes generated by the dual-path decoder. We performed copious experiments on three typical medical image segmentation tasks. The results show that our MC-DC network achieves better segmentation performance and lower computational complexity compared with the state-of-the-art segmentation network. Furthermore, the visualization of feature maps generated by encoder and decoder more intuitively illustrates the reasonableness of the designed components.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors greatly appreciate the financial supports of General Program National Natural Science Foundation of China (81971832).

References

[1] J. Chen, Y. Lu, Q Yu, et al., Transunet: Transformers make strong encoders for medical image segmentation, arXiv preprint (2021) arXiv:2102.04306.
[2] O. Ronneberger, P. Fischer, T: U-net Brox, Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 2015, pp. 234–241. October 5-9, 2015, Proceedings, Part III 18.
[3] Z. Zhou, MM. Rahman Siddiquee, N Tajbakhsh, et al., Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 2018, pp. 3–11. September 20, 2018, Proceedings 4.
[4] H. Huang, L. Lin, R Tong, et al., Unet 3+: A full-scale connected unet for medical image segmentation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 1055–1059.
[5] Ö. Çiçek, A. Abdulkadir, SS Lienkamp, et al., 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, 2016, pp. 424–432. October 17-21, 2016, Proceedings, Part II 19.
[6] 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), 2016, pp. 565–571.
[7] Y. Chen, K. Wang, X Liao, et al., Channel-Unet: a spatial channel-wise convolutional neural network for liver and tumors segmentation, Frontiers in genetics 10 (2019) 1110.
[8] Z. Fang, Y. Chen, D Nie, et al., Rca-u-net: Residual channel attention u-net for fast tissue quantification in magnetic resonance fingerprinting, Medical Image Computing and Computer Assisted Intervention–MICCAI 2019 (2019) 101–109.
[9] M. Noori, A. Bahri, Mohammadi, K: Attention-guided version of 2D UNet for automatic brain tumor segmentation, in: 2019 9th international conference on computer and knowledge engineering (ICCKE), 2019, pp. 269–275.
[10] B. Chen, Y. Liu, Z Zhang, et al., arXiv preprint, 2021.
[11] A. Vaswani, N. Shazeer, N Parmar, et al., Attention is all you need. Neural Information Processing Systems (NIPS), 2017, pp. 5998–6008.
[12] H. Cao, Y. Wang, J Chen, et al., Swin-unet: Unet-like pure transformer for medical image segmentation, in: Computer Vision–ECCV 2022 Workshops, Tel Aviv, Israel, 2023, pp. 205–218. October 23–27, 2022, Proceedings, Part III.
[13] Y. Zhang, H. Liu, Q Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 2021, pp. 14–24. September 27–October 1, 2021, Proceedings, Part I 24.
[14] JMJ. Valanarasu, P. Oza, I Hacihaliloglu, et al., Medical transformer: Gated axial-attention for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 2021, pp. 36–46. September 27–October 1, 2021, Proceedings, Part I 24.
[15] JMJ. Valanarasu, VM Patel, Unext: Mlp-based rapid medical image segmentation network, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, 2022, pp. 23–33. September 18–22, 2022, Proceedings, Part V.
[16] D. Lian, Z. Yu, X Sun, et al., As-mlp: An axial shifted mlp architecture for vision, arXiv preprint (2021) arXiv:2107.08391.
[17] K. Amara, A. Aouf, H Kennouche, et al., COVIR: A virtual rendering of a novel NN architecture O-Net for COVID-19 Ct-scan automatic lung lesions segmentation, Computers Graphics 104 (2022) 11–23.
[18] J. Hu, L. Shen, Sun, G: Squeeze-and-excitation networks, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.
[19] B. Yang, M. Liu, Y Wang, et al., Structure-guided segmentation for 3D neuron reconstruction, IEEE Trans Med Imaging 41 (4) (2021) 903–914.
[20] W. Shen, Y. Wang, M Liu, et al., Branch Aggregation Attention Network for Robotic Surgical Instrument Segmentation, IEEE Trans Med Imaging (2023).
[21] F. Yuan, Z. Zhang, Fang, Z: An effective CNN and Transformer complementary network for medical image segmentation, Pattern Recognition 136 (2023), 109228.
[22] K. He, X. Zhang, S Ren, et al., Deep residual learning for image recognition, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[23] Z. Liu, Y. Lin, Y Cao, et al., Swin transformer: Hierarchical vision transformer using shifted windows, arXiv preprint (2021) arXiv: 2103.14030.
[24] IO. Tolstikhin, N. Houlsby, A Kolesnikov, et al., Mlp-mixer: An all-mlp architecture for vision, Advances in Neural Information Processing Systems 34 (2021) 24261–24272.
[25] H. Touvron, P. Bojanowski, M Caron, et al., Resmlp: Feedforward networks for image classification with data-efficient training, IEEE Transactions on Pattern Analysis Machine Intelligence (2022).
[26] H. Liu, Z. Dai, D So, et al., Pay attention to mlps, Advances in Neural Information Processing Systems 34 (2021) 9204–9215.

[27] S. Chen, E. Xie, C Ge, et al., Cyclemlp: A mlp-like architecture for dense prediction, arXiv preprint (2021) arXiv:2107.10224.
[28] Y. Tang, K. Han, J Guo, et al., An image patch is a wave: Quantum inspired vision mlp, arXiv preprint (2021) arXiv:2111.12294.
[29] S-H. Gao, M-M. Cheng, K Zhao, et al., Res2net: A new multi-scale backbone architecture, IEEE transactions on pattern analysis machine intelligence 43 (2) (2019) 652–662.
[30] D-P. Fan, G-P. Ji, T Zhou, et al., Pranet: Parallel reverse attention network for polyp segmentation, Medical Image Computing and Computer Assisted Intervention–MICCAI 2020 (2020) 263–273.
[31] N. Codella, V. Rotemberg, P Tschandl, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), arXiv preprint (2019) arXiv:1902.03368.
[32] T. Mendonça, PM. Ferreira, JS Marques, et al., 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), 2013, pp. 5437–5440.
[33] D. Jha, PH. Smedsrud, MA Riegler, et al., Kvasir-seg: A segmented polyp dataset, in: MultiMedia Modeling: 26th International Conference, 2020, pp. 451–462.
[34] J. Silva, A. Histace, O Romain, et al., Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer, International journal of computer assisted radiology 9 (2014) 283–293.
[35] J. Zheng, H. Liu, Y Feng, et al., CASF-Net: Cross-attention and cross-scale fusion network for medical image segmentation, Computer Methods Programs in Biomedicine 229 (2023), 107307.
[36] W. Yang, A. Sirajuddin, X Zhang, et al., The role of imaging in 2019 novel coronavirus pneumonia (COVID-19), Eur Radiol 30 (2020) 4874–4882.
[37] M. Hosseiny, S. Kooraki, A Gholamrezanezhad, et al., Radiology perspective of coronavirus disease 2019 (COVID-19): lessons from severe acute respiratory syndrome and Middle East respiratory syndrome, Ajr Am J Roentgenol 214 (5) (2020) 1078–1082.
[38] I. Loshchilov, F Hutter, Sgdr: Stochastic gradient descent with warm restarts, arXiv preprint (2016) arXiv:1608.03983.
[39] J. Wang, L. Wei, L Wang, et al., Boundary-aware transformers for skin lesion segmentation, Medical Image Computing and Computer Assisted Intervention–MICCAI 2021 (2021) 206–216.
[40] O. Oktay, J. Schlemper, LL Folgoc, et al., Attention u-net: Learning where to look for the pancreas, arXiv preprint (2018) arXiv:1804.03999.
[41] Z. Zhang, Q. Liu, Y Wang, Road extraction by deep residual u-net, IEEE Geoscience Remote Sensing Letters 15 (5) (2018) 749–753.
[42] L-C. Chen, Y. Zhu, G Papandreou, et al., Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.
[43] Z. Gu, J. Cheng, H Fu, et al., Ce-net: Context encoder network for 2d medical image segmentation, IEEE Trans Med Imaging 38 (10) (2019) 2281–2292.
[44] R. Gu, G. Wang, T Song, et al., CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation, IEEE Trans Med Imaging 40 (2) (2020) 699–711.
[45] H. Wu, S. Chen, G Chen, et al., FAT-Net: Feature adaptive transformers for automated skin lesion segmentation, Med Image Anal 76 (2022), 102327.
[46] R. Azad, MT. Al-Antary, M Heidari, et al., Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model, IEEE Access 10 (2022) 108205–108215.
[47] H. Du, J. Wang, M Liu, et al., SwinPA-Net: Swin Transformer-based multiscale feature pyramid aggregation network for medical image segmentation, IEEE Transactions on Neural Networks Learning Systems (2022).
[48] C-H. Huang, H-Y. Wu, Y-L Lin, Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps, arXiv preprint (2021) arXiv:2101.07172.
[49] K. Patel, AM. Bur, G Wang, Enhanced u-net: A feature enhancement network for polyp segmentation, in: 2021 18th Conference on Robots and Vision (CRV). pp, 2021, pp. 181–188.
[50] A. Lin, B. Chen, J Xu, et al., Ds-transunet: Dual swin transformer u-net for medical image segmentation, IEEE Transactions on Instrumentation Measurement 71 (2022) 1–15.
[51] D. Jha, MA. Riegler, D Johansen, et al., Doubleu-net: A deep convolutional neural network for medical image segmentation, in: 2020 IEEE 33rd International symposium on computer-based medical systems (CBMS), 2020, pp. 558–564.
[52] H. Zhao, J. Shi, X Qi, et al., Pyramid scene parsing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.
[53] V. Badrinarayanan, A. Kendall, R: Segnet Cipolla, A deep convolutional encoder-decoder architecture for image segmentation, IEEE transactions on pattern analysis machine intelligence 39 (12) (2017) 2481–2495.
[54] W. Wang, E. Xie, X Li, et al., Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, arXiv preprint (2021) arXiv: 2102.12122.