frontiers | Frontiers in Neuroscience

ORIGINAL RESEARCH
published: 02 June 2022
doi: 10.3389/fnins.2022.876065

O-Net: A Novel Framework With Deep Fusion of CNN and Transformer for Simultaneous Segmentation and Classification

Tao Wang[1,2], Junlin Lan[1,2], Zixin Han[1,2], Ziwei Hu[1,2], Yuxiu Huang[1,2], Yanglin Deng[1,2], Hejun Zhang[3], Jianchao Wang[3], Musheng Chen[3], Haiyan Jiang[2,4], Ren-Guey Lee[5], Qinquan Gao[1,2,6], Ming Du[1], Tong Tong[1,2,6] and Gang Chen[3,7]**

[1] College of Physics and Information Engineering, Fuzhou University, Fuzhou, China, [2] Fujian Key Lab of Medical Instrumentation and Pharmaceutical Technology, Fuzhou University, Fuzhou, China, [3] Department of Pathology, Fujian Cancer Hospital, Fujian Medical University, Fuzhou, China, [4] College of Electrical Engineering and Automation, Fuzhou University, Fuzhou, China, [5] Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan, [6] Imperial Vision Technology, Fuzhou, China, [7] Fujian Provincial Key Laboratory of Translational Cancer Medicin, Fuzhou, China

OPEN ACCESS

Edited by:
Zhengwang Wu, University of North Carolina at Chapel Hill, United States

Reviewed by:
Yi Wang, Northwestern Polytechnical University, China
Anubha Gupta, Indraprastha Institute of Information Technology Delhi, India

***Correspondence:**
Tong Tong
ttraveltong@gmail.com
Gang Chen
naichengang@126.com

Specialty section:
This article was submitted to Original Research Article, a section of the journal Frontiers in Neuroscience

Received: *15 February 2022*
Accepted: *05 May 2022*
Published: *02 June 2022*

Citation:
Wang T, Lan J, Han Z, Hu Z, Huang Y, Deng Y, Zhang H, Wang J, Chen M, Jiang H, Lee R-G, Gao Q, Du M, Tong T and Chen G (2022) O-Net: A Novel Framework With Deep Fusion of CNN and Transformer for Simultaneous Segmentation and Classification. Front. Neurosci. 16:876065. doi: 10.3389/fnins.2022.876065

The application of deep learning in the medical field has continuously made huge breakthroughs in recent years. Based on convolutional neural network (CNN), the U-Net framework has become the benchmark of the medical image segmentation task. However, this framework cannot fully learn global information and remote semantic information. The transformer structure has been demonstrated to capture global information relatively better than the U-Net, but the ability to learn local information is not as good as CNN. Therefore, we propose a novel network referred to as the O-Net, which combines the advantages of CNN and transformer to fully use both the global and the local information for improving medical image segmentation and classification. In the encoder part of our proposed O-Net framework, we combine the CNN and the Swin Transformer to acquire both global and local contextual features. In the decoder part, the results of the Swin Transformer and the CNN blocks are fused to get the final results. We have evaluated the proposed network on the synapse multi-organ CT dataset and the ISIC 2017 challenge dataset for the segmentation task. The classification network is simultaneously trained by using the encoder weights of the segmentation network. The experimental results show that our proposed O-Net achieves superior segmentation performance than state-of-the-art approaches, and the segmentation results are beneficial for improving the accuracy of the classification task. The codes and models of this study are available at https://github.com/ortonwang/O-Net.

Keywords: CNN, transformer, medical image segmentation, deep learning, classification

1. INTRODUCTION

Image enhancement has been extensively performed on medical images based on morphology, such as clustering (Vasuda and Satheesh, 1713), edge detection (Patil and Deore, 2013), and threshold segmentation (Wang et al., 2015) to assist doctors in diagnosis in the early days. With the development of artificial intelligence, deep learning technology has been widely used in medical

image processing and analysis in recent years, and the accuracy of segmentation and classification on medical images is of great significance to the diagnosis of diseases today. In clinical practice, accurate image segmentation can provide clinicians with quantitative information, which can help clinicians make diagnostic decisions more precisely and efficiently (Liang et al., 2020). In addition, the additional information provided by computing methods is subjective and can avoid the objective bias by humans.

Nowadays, Convolutional Neural Network (CNN), especially Full Convolutional Network (FCN) is an effective segmentation method (Wang et al., 2021b) and it has been widely used in dense classification tasks such as semantic segmentation (Ji et al., 2020). Among different CNN networks, U-Net (Ronneberger et al., 2015) is a deep learning network with encoder and decoder structures, which has been widely used in medical image segmentation. In recent years, it has been widely used in medical image segmentation tasks due to its strong generalization. U-Net and its variants UNet++ (Zhou et al., 2018), UNet 3+ (Huang et al., 2020), CE-Net (Gu et al., 2019) have shown excellent performance in tasks, such as lesion segmentation, heart segmentation, and other organ segmentation. Based on the strong ability of learning and discriminating features, Res-UNet (Xiao et al., 2018) improves the performance of the network by introducing a residual network into the encoder part of U-Net. EfficientNet (Tan and Le, 2019) proposed a new scaling method that uniformly all dimensions of the depth, width, and resolution of the network through simple but efficient composite coefficients, which not only reduces a certain amount of calculation, but also improves the segmentation performance. Many experimental results have shown that the use of EfficientNet as an encoder can often further improve the performance of the network without increasing the amount of calculation.

However, these networks are faced with the common problem of CNN: it is difficult for CNN-based methods to learn the global and remote semantic information interaction (Chen et al., 2021) clearly. This is due to the fact that CNN extracts features with a convolutional process. Some studies tried to use image feature pyramid (Lin et al., 2017), atrous convolution layers (Chen et al., 2017, 2018; Gu et al., 2019), and self-attention mechanisms (Wang et al., 2018; Schlemper et al., 2019) to solve this problem. However, the global and remote semantic information is not fully learnt using these strategies. Inspired by the great success of transformer (Vaswani et al., 2017) in the field of natural language processing (NLP), researchers have tried to introduce transformer to make up for the shortcomings of CNN in global and remote information interaction. A transformer is an attention-based model and self-attention mechanism (SA) is a key component of transformer. It can model the correlation of all input tags which makes room for the transformer to deal with long-range dependencies. In Dosovitskiy et al. (2020), vision transformer (ViT) was applied to perform image recognition tasks and achieved relatively good results. After that, a novel framework called Swin Transformer (Liu et al., 2021) was proposed and significantly improved the performance of ViT in different tasks, such as image classification (Liu et al., 2021), object detection (Xu et al., 2021), and semantic segmentation (Xie et al., 2021). Based on the Swin Transformer, Cao et al. (2021) proposed Swin-Unet, which combined the U-Net structure and Swin Transformer for medical image segmentation, the encoding part and the decoding part in Swin-Unet were both performed using Swin Transformer. With the proposal of these methods, the accuracy of segmentation tasks is further improved. However, the input in transformer is formed as one-dimensional sequence. The transformer networks focus on learning the global contextual information, but may lose some local details. Therefore, it is beneficial to combine the global information learnt by transformer and the local information by CNN to enrich the learnt features.

Based on the advantages of CNN and transformer, we propose an O-Net framework to combine the CNN and the transformer to learn both global and local contextual features. We combine the CNN and Swin Transformer as encoder first and send them into a CNN-based decoder and a Swin Transformer-based decoder, respectively. The results of two decoders are fused to get the final result. This network combines the advantages of CNN and transformer and may improve the performance of medical image segmentation. Our experimental results have shown that the performance of the network can be significantly improved by combining CNN and transformer. In addition, a classification task is simultaneously performed based on the O-Net. Experiments show that the segmentation results are beneficial for improving the accuracy of the classification task. Experiments on the synapse multi-organ segmentation dataset and the ISIC2017 skin lesion challenge dataset have demonstrated the superiority of our method compared to other state-of-the-art segmentation methods. In addition, based on the segmentation network, the performance of the classification network has also been greatly improved.

2. RELATED WORKS

CNN-based methods: CNN is a kind of feedforward neural network that includes convolution calculations and has a deep structure. It is one of the representative algorithms of deep learning. Lenet[18] first defined the CNN network structure in 1998, and it was not until the publication of AlexNet (Krizhevsky et al., 2012) in 2012 that CNN has gradually become mainstream. Since then, lots of efficient and deep convolutional neural networks have been proposed. For example, VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), DenseNet (Huang et al., 2017), GoogleNet (Szegedy et al., 2015), HRNet (Sun et al., 2019), Inception v3 (Szegedy et al., 2016), and EfficientNet (Tan and Le, 2019). These networks perform well in various applications. In addition to these network innovations, new convolutional layers such as deformable convolution (Dai et al., 2017; Zhu et al., 2019) and depth-wise convolution (Xie et al., 2017) were proposed for different tasks. With the development of CNN, U-Net was proposed and widely used in segmentation tasks because of its simple structure, good effects, and strong generalization. After that, various U-shape network based U-Net

have been proposed such as U-SegNet (Kumar et al., 2018), Res-UNet (Xiao et al., 2018), Dense-UNet (Li et al., 2018), U-Net++ (Zhou et al., 2018), U-2-Net (Qin et al., 2020), and UNet3+ (Huang et al., 2020) CE-Net (Gu et al., 2019). Gehlot et al. (2020) proposed an Encoder-Decoder based CNN with Nested-Feature Concatenation (EDNFC-Net) for automatic segmentation. Some networks introduce novel structures in the encoder part while others in the decoder part. Because of the strong generalization of the network, the U-shaped architecture network has also been extended to 3D medical image segmentation, such as 3D-UNet (Çiçek et al., 2016) and V-Net (Milletari et al., 2016). Moreover, Gehlot et al. proposed AION (Gehlot and Gupta, 2021), an architecture with two coupled networks and classification heads which is applicable for stain normalization, classification, and segmentation tasks.

Transformers: Transformer was first proposed for machine translation and achieved the best performance in many NLP tasks. To combine computer vision (CV) and natural language processing (NLP) domain knowledge, researchers developed Vision Transformer (ViT) (Dosovitskiy et al., 2020) by directly applying transformers with global self-focus to full-size images. The ViT model achieved both high efficiency and accuracy in image recognition tasks. Based on ViT, Chen et al. (2021) proposed the first transformer-based medical image segmentation framework TransUNet which further improved the accuracy of image segmentation tasks. However, ViT needs to be pre-trained on its large datasets to achieve good performance. To solve this problem, some training schemes were designed in Deit (Touvron et al., 2021) so that the algorithm can perform well on smaller data sets. To further improve the accuracy, a new vision transformer called Swin Transformer (Liu et al., 2021) was proposed, it is a hierarchical transformer whose representation is computed with Shifted windows. This hierarchical architecture has the flexibility of modeling at various scales and has linear computational complexity relative to the image size. These features make it compatible with many vision tasks, including image classification and semantic segmentation. Based on Swin Transformer, Cao et al. (2021) proposed a pure transformer U-shaped encoder-decoder network named Swin-Unet for medical image segmentation, which has relatively good performance in some datasets.

Self-attention/transformer combined with CNN: In recent years, researchers have tried to improve the performance of the network through the self-attention mechanism (Wang et al., 2018) to overcome the shortcomings of CNN learning global semantic information. In Schlemper et al. (2019), the skip-connections with additive attention gate were integrated with U-shaped architecture to improve medical image segmentation. But this is still the method based on CNN after all and it has not completely solved the limitation of learning global information. Several studies have been carried out to combine CNN and transformer. TransUNet (Chen et al., 2021) was proposed by combining the advantages of transformer and CNN. The transformer encodes image patches from a CNN feature map as the input sequence for extracting global contexts. A mixed transformer module (MTM) (Wang et al., 2021a) was proposed for simultaneous inter- and intra- affinities learning. TransFuse (Zhang et al., 2021) combines transformers and CNNs in a parallel style to capture both global dependency and low-level spatial details efficiently in a much shallower manner for medical image segmentations. Liang et al. (2022) proposed transconver with a parallel module named transformer-convolution inception which extracts local and global information *via* convolution blocks and transformer blocks, respectively. TransMed (Dai et al., 2021) was proposed for multi-modal medical image classification which combines the advantages of CNN and transformer to extract low-level features of images efficiently and establish long-range dependencies between modalities. These algorithms improve the global attention of the model based on their complementarity by directly combining CNN and transformer.

3. THE PROPOSED METHOD

3.1. Overall Architecture Design

A schematic view of the proposed O-Net is presented in **Figure 1**. O-Net is composed of two parts: an encoder module and a decoder module. The basic units of O-Net include the Swin Transformer block, EfficientNet block, and CNN Decoder block. During the segmentation task, the encoder module extracts the features of the input image to obtain the high-dimensional and low-dimensional features, which are then decoded back to the full spatial resolution by the decoder module. After extracting the features in the encoder part, the segmentation network provided an interface to integrate a classification network for simultaneously performing the classification task. Each module is described in detail below.

3.2. Swin Transformer Block

Different from the transformer, Swin Transformer is built based on shifted windows rather than the standard multi-head self attention (MSA) module. Two consecutive Swin Transformer blocks are presented in **Figure 2**. Each Swin Transformer block consists of residual connection and 2-layer MLP with Gaussian Error Linear Units (GELU) non-linearity, LayerNorm (LN) layer, and multi-head self attention module. The shifted window-based multi-head self attention (SW-MSA) module and the window-based multi-head self attention (W-MSA) module are applied in the two successive transformer blocks, respectively. Based on such a window partitioning approach, successive Swin Transformer blocks can be formulated as follows:

$$\hat{z}^l = W - MSA(LN(z^{l-1})) + z^{l-1}, \tag{1}$$

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l, \tag{2}$$

$$z^{l+1} = SW - MSA(LN(z^l)) + z^l, \tag{3}$$

$$z^{l+1} = MLP(LN(z^{l+1})) + z^{l+1}, \tag{4}$$

Where z^l and $\hat{z}^l$ represent the output features of the (S)W-MSA module and the MLP module of the l^{th} block, respectively. Similar

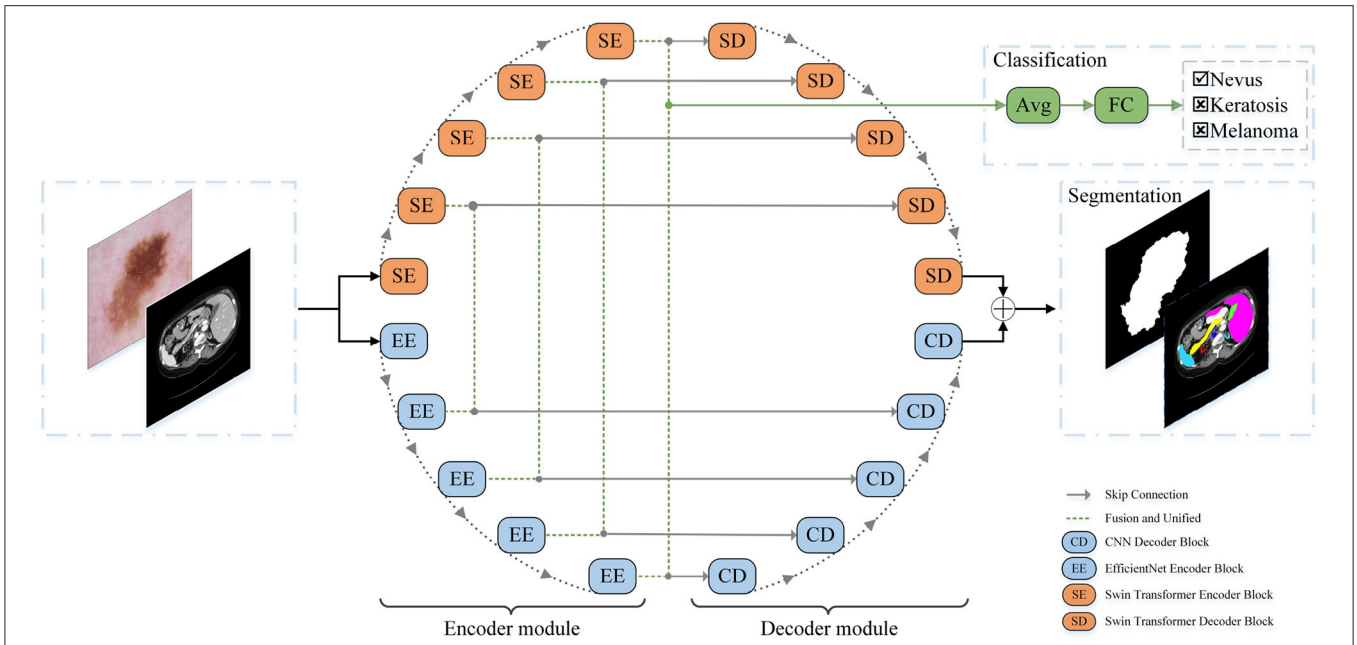

FIGURE 1 | The architecture of our proposed O-Net.

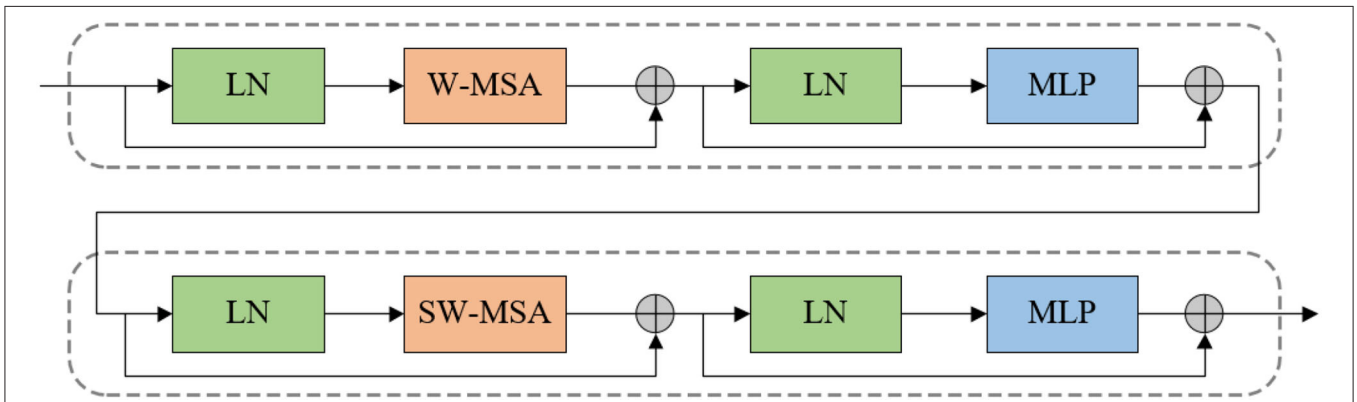

FIGURE 2 | Two successive Swin Transformer block.

to the previous works (Hu et al., 2018, 2019), self-attention is computed as follows:

$$Attention(Q, K, V) = SoftMax(\frac{QK^T}{\sqrt{d}} + B)V, \quad (5)$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ denote the query, key, and value matrices. M^2 represents the number of patches in a window, and d is the query dimension. Since the relative position along each axis is within the range[−M+1, M−1], the values in B are taken from the bias matrix $\hat{B} \in \mathbb{R}^{(2M-1)\times 2M+1}$.

3.3. EfficientNet Block

EfficientNet block (Tan and Le, 2019) was proposed based on a neural structure search. This block uses composite coefficients to uniformly scale the depth, width, and resolution of the network. A schematic view of the EfficientNet block is presented in **Figure 3**. Each EfficientNet block is composed of MBConvBlocks (Sandler et al., 2018) which consists of convolution, batch normalization, and Swish activation layers. The network achieves better performance with the same parameters by uniformly scaling the network width, depth, or resolution in a fixed proportion. We employ the EfficientNet block as the encoder part of CNN to extract features efficiently and effectively.

3.4. Encoder Module

In the encoder part, we combine EfficientNet and Swin Transformer. For the Swin Transformer Encoder, it is composed of Swin Transformer Block and patch merging layer. Images are separated into non-overlapping patches with a patch size of

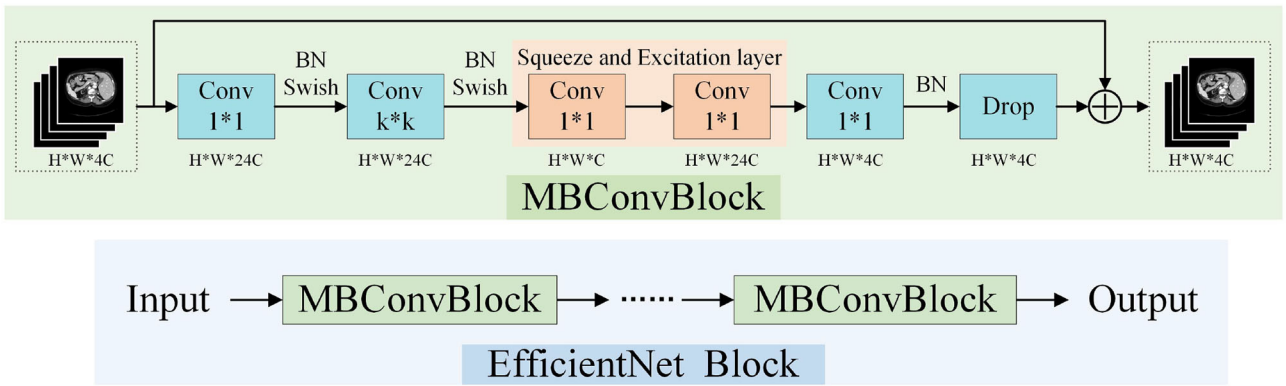

FIGURE 3 | The architecture of EfficientNet Block.

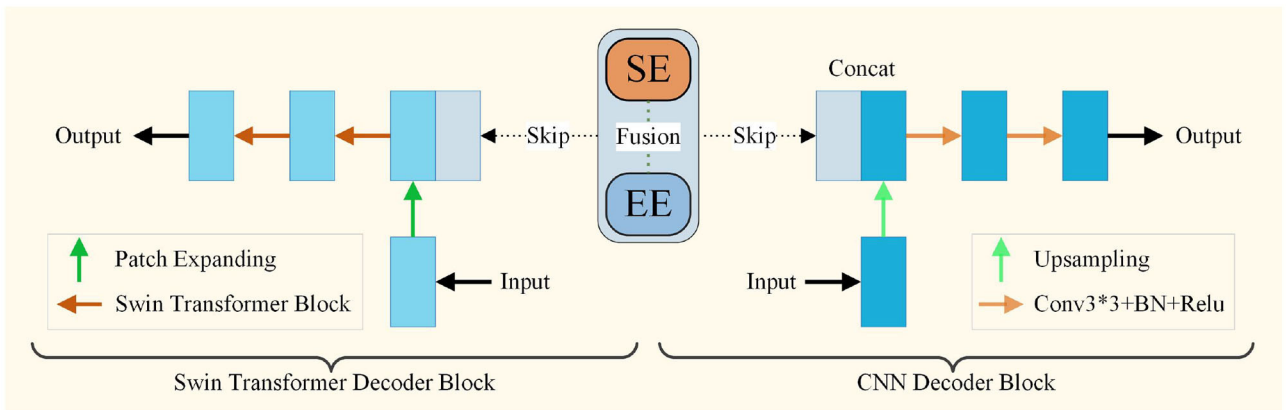

FIGURE 4 | The architecture of decoder module.

4×4 to transform the inputs into sequence embeddings, then concatenated together by the patch merging layer. The feature resolution will be down-sampled by 2× after such processing, and the feature dimension of each patch becomes to 4×4×3 = 48. Furthermore, a linear embedding layer is applied to project feature dimension into an arbitrary dimension (represented as C). The transformed patch tokens pass through several Swin Transformer blocks and patch merging layers to generate the hierarchical feature representations.

For the EfficientNet encoder, the input image is convoluted and down-sampled first. Feature extraction is carried out through the EfficientNet block which uniformly scales the depth, width, and resolution of the network through composite coefficients. We can achieve relatively efficient feature extraction with only a small amount of computation using this module. Since the feature dimensions of two encoders are different, it is required to normalize the dimension before fusing them. The features extracted by the Swin Transformer are set to C×H×W using a linear projection. After that, the features are fused with the features extracted by the EfficientNet block *via* skip-connections. Similarly, when using the Swin Transformer decoder, we project the features extracted by the EfficientNet block through the linear embedding layer and fuse them with the features extracted by the Swin Transformer encoder.

3.5. Decoder Module

The decoder module is adopted to restore the high-level semantic features extracted from the encoder module. The decoder part consists of the Swin Transformer decoder block and the CNN decoder block. A schematic view of the decoder modules is presented in **Figure 4**. The Swin decoder block is composed of a patch expanding layer and a Swin Transformer block. The features extracted by the encoder are multi-scale fused through skip-connections. The patch expanding layer reshapes feature maps of adjacent dimensions into large feature maps with 2× up-sampling of resolution. In the end, the last patch expanding layer is used to perform 4× up-sampling to restore the resolution of the feature maps to the input resolution (W×H), and a linear projection layer is applied on these up-sampled features to output the pixel-level segmentation predictions.

The CNN decoder block is composed of a 2× upsampling operator, two 3×3 convolution layers, and a batch normalization layer with a Rectified Linear Units(ReLU) layer. Simple

upsampling and convolution are two common operations of the decoder in the CNN decoder blocks. After the 2× upsampling operator, the features were fused with those from encoders through skip-connection. After the two convolution processes, the features are input for the next decoder. At the end of the decoder, a convolution layer is applied to output the pixel-level segmentation predictions. Finally, the outputs of the two decoders are fused to obtain the final segmentation result.

3.6. Classification Method

The encoder part of the segmentation network and the classification network share the same structure. When encoders perform the classification task, the role of encoders is to extract contextual features and locate the target region like the segmentation task. The primary task of classification aims to accurately locate the target area, and the purpose of the segmentation network is to realize it. Therefore, after the training of the segmentation network, we use the learned weights of the encoder in the network as the initial parameters of the classification network. After that, we utilize the features of the lowest dimension in the encoder through the average pooling layer and a fully connected layer (FC) to perform the classification task.

4. EXPERIMENTS

4.1. Datasets

Synapse multi-organ segmentation dataset (synapse): The dataset includes 30 abdominal CT scans from MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge. Each CT volume consists of 85−198 slices of 512×512 pixels and there are 3,779 axial abdominal clinical CT images in total. Following Chen et al. (2021) and Liu et al. (2021), 18 samples were used as the training set and 12 samples as the testing set. The annotation of each image includes 8 abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, and stomach). The dice metric and the average Hausdorff Distance (HD) are used to evaluate our method on this dataset. The dice metric evaluates the degree of pixel overlap between the ground truth and prediction results and it is calculated as follows:

$$Dice = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{6}$$

where TP, FP, and FN refer to the number of true positives, false positives, and false negatives, respectively, besides, TN means true negatives. The HD calculates the maximum distance between the contours of the ground truth and predicted results, which can be formulated as follows:

$$H(A, B) = max(h(A, B), h(B, A) \tag{7}$$

$$h(A, B) = \max_{a \in A} \left\{ \min_{b \in B} \{ \|a - b\| \} \right\} \tag{8}$$

$$h(B, A) = \max_{b \in B} \left\{ \min_{a \in A} \{ \|b - a\| \} \right\} \tag{9}$$

where A and B denote the contours of the ground truth and predicted results, respectively, and h(A,B) denotes the unidirectional HD from A to B.

ISIC2017 skin lesion challenge dataset (ISIC2017): The 2017 International Skin Imaging Collaboration (ISIC) skin lesion segmentation challenge dataset (Codella et al., 2018) includes 2,000 training images, 150 validation images, and 600 test dermoscopic images. Each image is paired with an expert manual tracing of skin lesion boundaries for the segmentation task and the lesion gold standard diagnosis (i.e., nevus, melanoma, and seborrheic keratosis) for the classification task. The size of the images in the dataset varies from 453×679 to 4499×6748 pixels. We used Dice, Mean Intersection over Union (IoU), Precision (Pre), Recall, F1-score, and Pixel Accuracy (PA) as the metrics to evaluate the accuracy of the segmentation work. In addition, we used Accuracy (AC), F1-score, precision (Pre), and specificity (SP) as the metrics to evaluate the classification task. These metric are calculated as follows:

$$IoU = \frac{TP}{TP + FN + FP} \tag{10}$$

$$Pre = \frac{TP}{TP + FP} \tag{11}$$

$$PA = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$AC = \frac{TP + TF}{TP + TN + FP + FN} \tag{13}$$

$$F1 - score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{14}$$

4.2. Implementation Details

Our method was implemented based on the Pytorch Deep Learning framework using python. For all training cases, flips and rotations were used as data augmentation to improve the generalization ability of the model. We trained our model on an Nvidia RTX 3090 GPU with 24GB memory. The input image size was set to 224×224 on the synapse dataset and 512×512 on the ISIC2017 dataset. The patch on the size was set to 4 in both tasks. All encoders and Swin Transformer blocks in the model were pretrained on ImageNet (Deng et al., 2009). During the training process of the synapse dataset, the batch size was set to 24 and the popular SGD optimizer with momentum of 0.9 and weight decay of 1e-4 and a learning rate of 1e-4 is used for the backpropagation of the model. During the process of ISIC2017 dataset, the models were optimized by AdamW with a learning rate of 1e-4 and a batch size of 8.

4.3. Experiment Results on the Synapse Dataset

The comparison of the proposed O-Net with previous state-of-the-art methods on the synapse multi-organ CT dataset is presented in **Table 1**. Experimental results demonstrate that our

TABLE 1 | Experimental results of different methods on the synapse multi-organ CT dataset.

Method	Dice↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
V-Net Milletari et al. (2016)	68.81	–	75.34	51.87	77.10	**80.75**	87.84	40.05	80.56	56.98
DARR Fu et al. (2020)	69.77	–	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 ViT Chen et al. (2021)	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
U-SegNet Kumar et al. (2018)	72.61	43.94	85.69	64.33	75.12	66.41	91.72	50.59	84.07	62.96
R50 U-Net Chen et al. (2021)	74.68	36.87	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
AION Gehlot and Gupta (2021)	75.54	32.27	87.59	58.74	82.47	73.45	93.47	49.44	87.52	71.61
R50 Att-UNet Chen et al. (2021)	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
U-Net Ronneberger et al. (2015)	76.85	39.7	89.07	**69.72**	77.77	68.60	93.43	53.98	86.67	75.58
EDNFC-Net Gehlot et al. (2020)	77.21	35.07	86.08	62.47	84.31	78.27	92.61	57.31	85.36	71.24
TransUNet Chen et al. (2021)	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Att-UNet Oktay et al. (2018)	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransFuse Zhang et al. (2021)	78.95	26.59	87.09	61.64	82.20	76.91	94.19	59.01	89.86	80.73
Swin-Unet Cao et al. (2021)	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	**90.66**	76.60
O-Net	**80.61**	**21.04**	88.36	67.45	**84.44**	77.13	**95.24**	**61.52**	90.03	**80.74**

The symbol ↑ means the higher value, the better.
The symbol ↓ means the lower value, the better.
Bold font to highlight the optimal values.

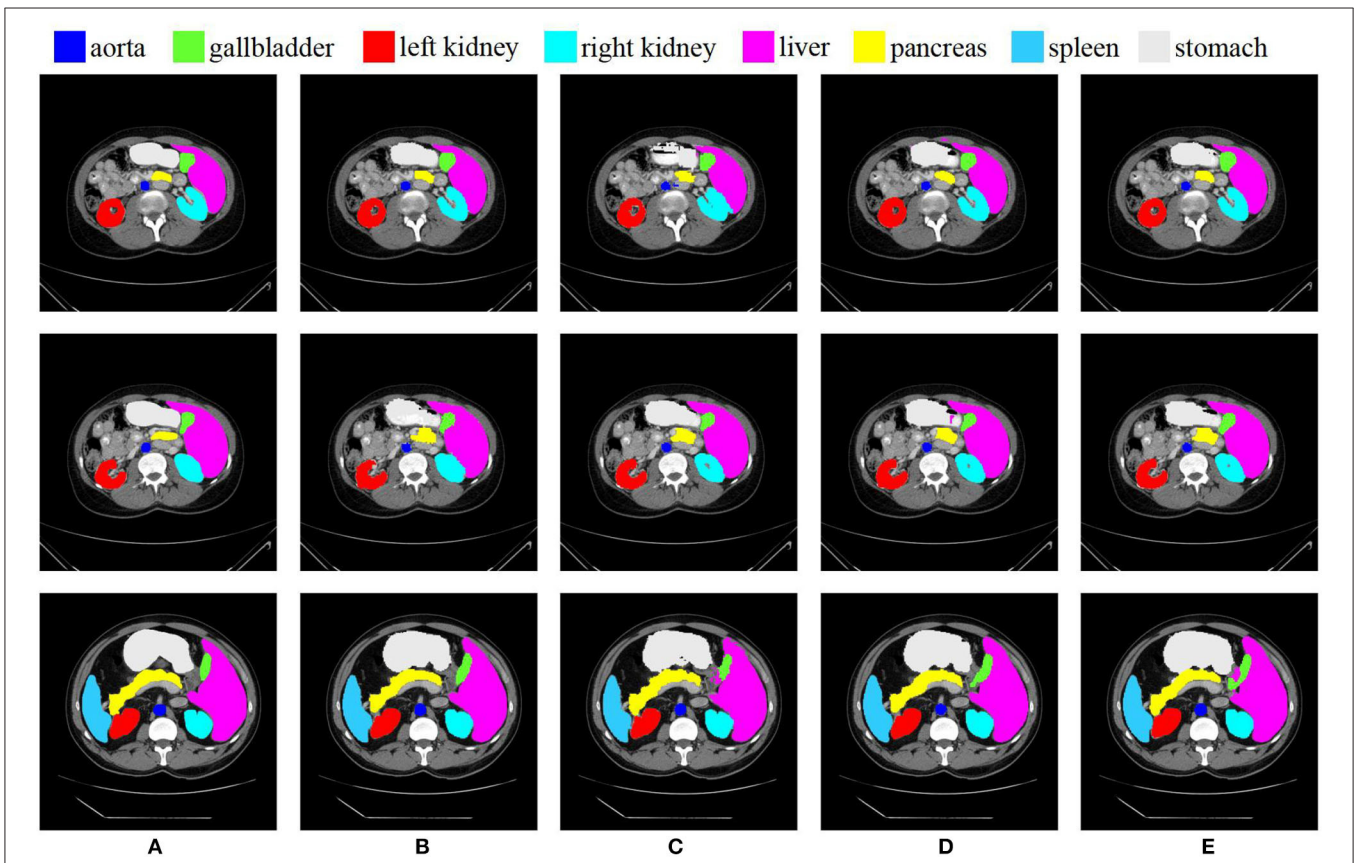

FIGURE 5 | Conparision of different methods on the Synapse multi-organ dataset by visualization. From left to right: **(A)** Ground Truth, **(B)** O-Net, **(C)** SwinUNet, **(D)** TransUNet, and **(E)** R50 AttUNet.

algorithm achieves the best performance with a segmentation accuracy of 80.61% (Dice↑) and 21.04 (HD↓) performance. We can see from the results that the CNN-based method performs worse on edge predictions than the transformer method from the metric of HD. This also indicates that our algorithm not only performs better in terms of segmentation, but also has

TABLE 2 | Segmentation results of different methods on the ISIC2017 dataset.

Method	Dice	mIoU	Pre	recall	F1-score	PA
U-Net Ronneberger et al. (2015)	85.22	78.40	91.17	73.98	77.80	91.19
R50-U-Net Xiao et al. (2018)	87.48	80.86	92.99	78.19	81.70	92.19
U-SegNet Kumar et al. (2018)	87.87	81.22	90.50	81.13	82.49	92.33
ENDFC-Net Gehlot et al. (2020)	88.00	81.43	90.26	**82.10**	82.80	92.29
M-Net Fu et al. (2018)	88.33	82.25	94.46	79.04	83.38	92.67
AION Gehlot and Gupta (2021)	88.84	82.56	92.26	81.95	84.02	92.88
CE-Net Gu et al. (2019)	89.64	83.56	95.40	80.47	84.99	93.67
Swin-Unet Cao et al. (2021)	88.77	82.69	94.64	79.16	83.51	94.04
TransFuse Zhang et al. (2021)	89.63	83.78	95.56	80.35	84.75	93.73
TransUNet Chen et al. (2021)	89.99	84.21	95.59	81.21	85.42	93.97
O-Net	**90.30**	**84.52**	**95.65**	81.72	**85.89**	**94.09**

Bold font to highlight the optimal values.

TABLE 3 | Classification accuracy of different methods on the ISIC2017 dataset.

	Average	Nevus classification			
Method	**AC**	**AC**	**F1-score**	**Pre**	**SP**
Swin Transformer Liu et al. (2021)	80.22	89.50	81.18	62.16	91.76
AION Gehlot and Gupta (2021)	81.55	85.33	76.01	50.74	86.86
TransMed Dai et al. (2021)	84.11	89.19	80.10	61.90	92.16
MobileNetV3 Howard et al. (2019)	84.89	89.33	81.53	60.83	90.78
EfficientNet-B3 Tan and Le (2019)	85.22	90.67	82.64	66.67	93.33
Inception v4 Szegedy et al. (2016)	85.33	89.16	81.45	60.16	90.39
ResNet50 He et al. (2016)	85.44	91.00	82.97	68.37	93.92
DenseNet201 Huang et al. (2017)	86.56	**92.00**	**85.36**	69.81	93.73
O-Net	**87.22**	91.67	83.51	**72.73**	**95.29**

	Average	Melanoma classification				Keratosis classification			
Method	**AC**	**AC**	**F1-score**	**Pre**	**SP**	**AC**	**F1-score**	**Pre**	**SP**
Swin Transformer Liu et al. (2021)	80.22	73.00	71.63	84.07	73.91	78.17	68.45	45.33	83.02
AION Gehlot and Gupta (2021)	81.55	77.33	75.69	85.40	74.40	82.00	69.73	54.46	90.48
TransMed Dai et al. (2021)	84.11	79.17	77.13	84.72	71.50	84.00	79.83	59.63	55.56
MobileNetV3 Howard et al. (2019)	84.89	80.50	78.78	86.51	75.36	84.83	74.59	62.75	92.13
EfficientNet-B3 Tan and Le (2019)	85.22	80.50	78.82	86.70	75.85	84.50	**75.71**	59.84	89.86
Inception v4 Szegedy et al. (2016)	85.33	80.83	79.05	86.39	74.88	86.00	75.94	67.37	93.58
ResNet50 He et al. (2016)	85.44	81.50	79.99	**87.90**	**78.26**	83.83	75.28	57.69	88.61
DenseNet201 Huang et al. (2017)	86.56	83.50	81.55	86.57	73.91	84.17	72.48	61.96	92.75
O-Net	**87.22**	**84.17**	**81.58**	84.49	67.63	**85.83**	74.19	**70.00**	**95.03**

Bold font to highlight the optimal values.

a good performance in edge prediction. For organs with high segmentation difficulty such as Pancreas and Gallbladder, our method obtains the best and the third results, respectively, which also reflects the strong generalization of our algorithm. The specific segmentation results of different algorithms on this dataset are presented in **Figure 5**. In this work, we demonstrate that the in-depth combination of CNN and Swin Transformer can learn both the global and the local contextual features, thereby obtaining better segmentation results.

4.4. Experiment Results on the ISIC2017 Dataset

We further evaluated the proposed method for medical image segmentation and classification using the ISIC2017 dataset. The results of segmentation and classification are presented in **Tables 2**, **3**. From **Table 2**, we can see that in the segmentation task, the combination of the CNN and the Swin Transformer can achieve better performance than that of single CNN or that of only the Swin Transformer. This indicates the effectiveness of the

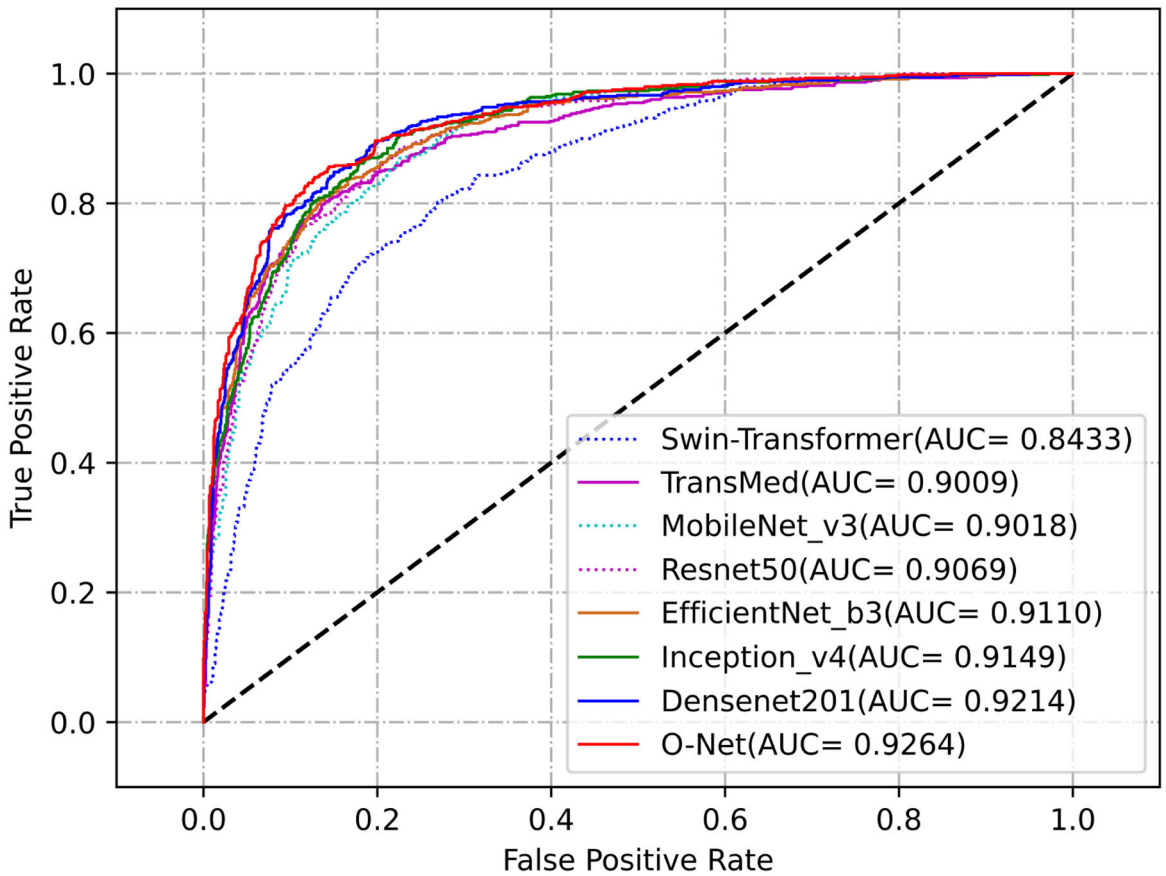

FIGURE 6 | Receiver Operating Characteristic curves of the different methods for classification task on the ISIC2017 dataset.

combination of these two structures. The O-Net has achieved the best performance in the six metrics which reflects the superiority of our method. The Receiver Operating Characteristic (ROC) curves of the classification methods are shown in **Figure 6**. The Area Under Curve (AUC) value for O-Net is 0.9264 which is the best performance among compared methods. Based on the data from **Table 3** and the ROC curves of the classification task, we can see that O-Net has also achieved excellent performance in the classification task. The specific segmentation results of different algorithms on this dataset are presented in **Figure 7**. The experimental results of classification tasks on this dataset indicate that combining CNN and Swin Transformer for classification tasks can improve the accuracy of the classification tasks. The performance can be further improved by initializing the classification network with the parameters from the encoder part of the segmentation network.

4.5. Ablation Study

The results of the ablation studies are shown in **Tables 4**, **5**. We will compare and analyze the effects of different factors on the segmentation performance in the following sections.

Effect of encoder: The experimental results in **Table 4** show that the best results are achieved by using the EfficientNet block as the encoder, while the number of parameters is not large. The parameter quantity of the MobileNet is smaller than that of the EfficientNet, but its accuracy is far too poor than the others. The accuracy of Inception v3 is similar to ours, but the amount of calculation is much larger than that of EfficientNet. Therefore, we use EfficientNet as a CNN-based encoder.

Effect of combination: The segmentation network consists of encoder and decoder. How to combine the CNN based method and the Swin Transformer based method is a point worth exploring. **Table 5** shows the effects of adopting different models for encoder and decoder. It can be seen from the results that the best performance is achieved by combining them in both the encoder and decoder parts. As can be seen from the results, better segmentation performance is achieved when CNN is used in the encoder part and Swin Transformer is used in the decoder part.

Effect of learning rate and batch size: To explore the best learning rate and batch size in the training process of the algorithm, we carried out a series of experiments. The experimental results are shown in **Table 6**. It can be seen from the top half of the chart that the best Dice was obtained when the learning rate was set to 1e-2. Although the best HD was obtained when the learning rate was set to 1e-1, its Dice was lower, therefore, we chose the learning rate of 5e-2. We can

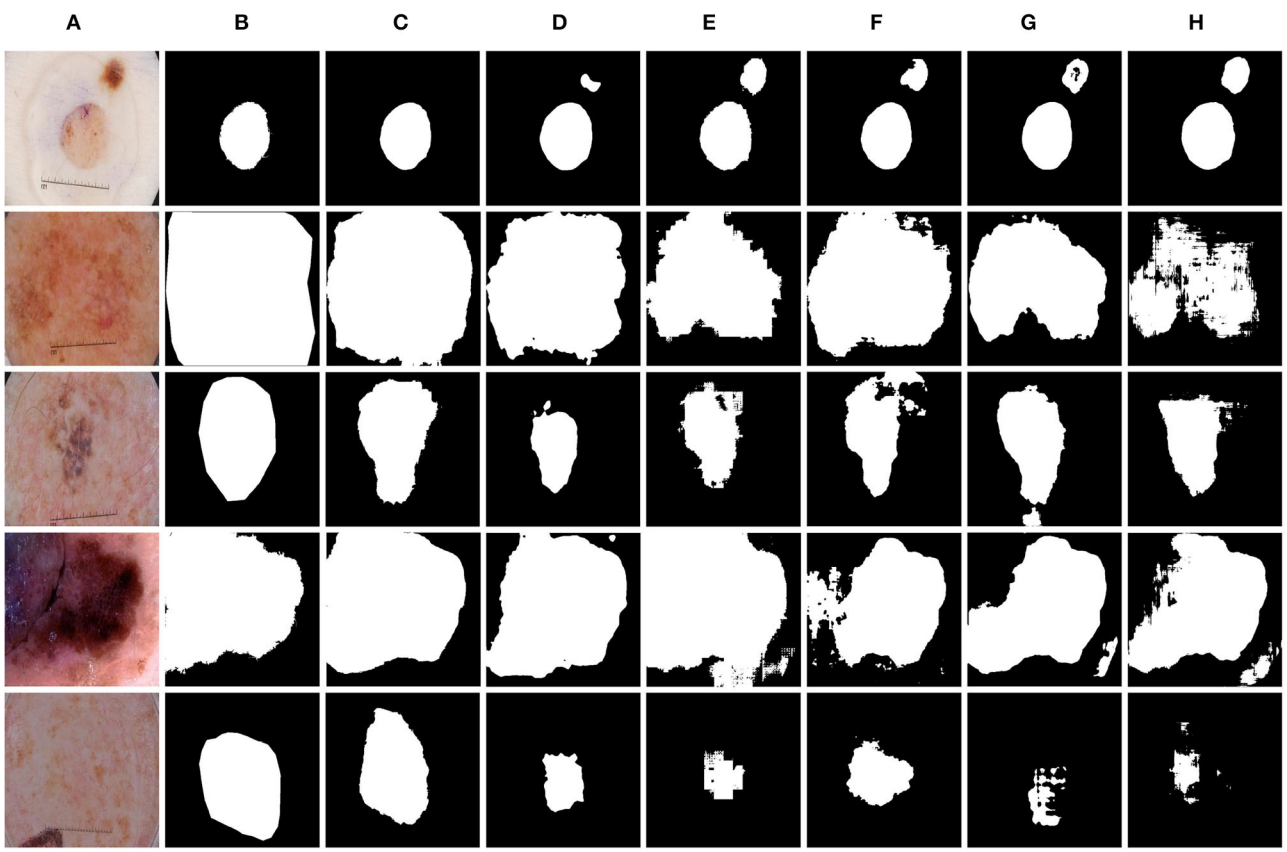

FIGURE 7 | Comparison of different methods on the ISIC2017 dataset by visualization. **(A)** Image, **(B)** Ground Truth, **(C)** O-Net, **(D)** TransUNet, **(E)** Swin-UNet, **(F)** CE-Net, **(G)** R50 AttUNet, and **(H)** UNet.

TABLE 4 | Ablation study on the encoder of CNN method.

Encoder	Params	Dice↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
MobileNetV3 Howard et al. (2019)	**5.48**	76.66	26.27	86.12	62.25	82.07	**90.65**	94.06	55.72	88.62	73.77
DenseNet201 Huang et al. (2017)	20.01	78.91	20.45	87.52	65.52	82.61	78.20	95.05	57.42	86.40	78.65
Resnet50 He et al. (2016)	25.55	79.16	23.01	87.71	66.86	81.73	75.22	94.18	58.86	**90.42**	78.31
Inception v3 Szegedy et al. (2016)	23.83	80.36	22.78	88.09	63.76	82.19	79.25	95.16	**65.17**	87.12	**82.13**
EfficientNet-b3 Tan and Le (2019)	12.23	**80.61**	**21.04**	**88.36**	**67.45**	**84.44**	77.13	**95.24**	61.52	90.03	80.74

The symbol ↑ means the higher value, the better.
The symbol ↓ means the lower value, the better.
Bold font to highlight the optimal values.

also draw from the bottom half of the chart that the best dice was obtained when the batch size was set to 24. Although the HD is lower when batch size was set to 8 and 6, the Dice of the Gallbladder is far too low, which is not conducive to the overall segmentation, therefore, the batch size of 24 would be more appropriate.

5. CONCLUSION

We introduce a novel method based on the combination of CNN and Swin Transformer for medical image segmentation and classification. To make full use of the global and the local information to improve medical image segmentation and classification, we propose O-Net, which combines the advantages of these two structures for improving both the segmentation and the classification performance. We combine CNN and transformer in both encoder and decoder parts of the network. In addition, we have shown that the proposed segmentation network is beneficial for the classification task. Experimental results have demonstrated that the proposed O-Net achieves competitive performance and good generalization ability in both the segmentation and the classification tasks.

TABLE 5 | Ablation study on the combination of CNN method and Swin Transformer method.

Encoder		Decoder											
Efficient net-block	Swin transformer	CNN decoder	Swin transformer decoder	Dice↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
✓		✓		78.86	28.86	87.72	62.19	83.11	76.67	94.49	56.61	89.48	80.58
✓			✓	79.93	26.88	87.90	68.09	83.89	76.05	94.42	**62.95**	87.32	78.86
✓		✓	✓	80.34	22.53	**88.67**	67.38	83.95	77.01	95.12	60.06	88.76	**81.77**
	✓	✓		77.55	31.03	86.14	63.49	81.59	75.82	93.68	54.61	90.19	74.87
	✓		✓	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	**90.66**	76.60
	✓	✓	✓	79.38	22.34	87.60	62.53	84.86	**80.54**	94.42	58.75	90.64	75.66
✓	✓	✓		79.47	29.19	87.71	66.21	81.64	74.69	94.65	61.61	89.19	80.02
✓	✓		✓	80.41	27.33	86.74	71.19	84.32	77.29	94.30	60.63	89.2	79.64
✓	✓	✓	✓	**80.61**	**21.04**	88.36	**67.45**	**84.44**	77.13	**95.24**	61.52	90.03	80.74

The symbol ↑ means the higher value, the better.
The symbol ↓ means the lower value, the better.
Bold font to highlight the optimal values.

TABLE 6 | Ablation study on learning rate and batch size.

Learn rate	Dice↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
1e-1	79.21	**20.06**	86.56	63.48	**84.61**	**77.14**	94.32	56.99	**91.90**	78.69
5e-2	**80.61**	21.04	**88.36**	67.45	84.44	77.13	**95.24**	**61.52**	90.03	80.74
1e-2	79.07	20.14	87.64	67.74	81.95	74.69	94.71	58.33	89.44	78.03
5e-3	79.76	23.07	88.18	**68.51**	83.60	76.92	94.42	58.84	88.59	79.03
1e-3	76.57	30.37	85.28	62.96	81.61	74.51	92.96	54.13	86.37	**84.70**
Batch size	**Dice↑**	**HD↓**	**Aorta**	**Gallbladder**	**Kidney(L)**	**Kidney(R)**	**Liver**	**Pancreas**	**Spleen**	**Stomach**
8	78.81	**15.67**	88.12	44.45	**84.59**	**80.24**	94.73	**67.40**	89.97	81.00
16	78.36	18.25	**88.69**	38.12	84.57	79.46	95.16	66.20	**91.44**	**83.27**
24	**80.61**	21.04	88.36	**67.45**	84.44	77.13	95.24	61.52	90.03	80.74
32	80.35	27.93	88.32	66.70	81.94	76.19	**95.31**	64.06	88.94	81.27

The symbol ↑ means the higher value, the better.
The symbol ↓ means the lower value, the better.
Bold font to highlight the optimal values.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

TW, JL, ZHa, ZHu, YH, YD, QG, MD, TT, and GC: concept and design. TW, JL, HZ, JW, MC, and TT: acquisition of data. TW, JL, ZHa, ZHu, QG, and TT: model design. TW, JL, ZHa, ZHu, YH, YD, and TT: data analysis. TW, JL, ZHa, ZHu, YH, YD, TT, and GC: manuscript drafting. TW, JL, ZHa, ZHu, YH, YD, HZ, JW, MC, HJ, R-GL, QG, MD, TT, and GC: approval. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61901120 and 62171133, the Science and Technology Program of Fujian Province of China under Grant No. 2019YZ016006, and Health and Family Planning Research Talent Training Program of Fujian Province under Grant No. 2020GGB009.

REFERENCES

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2021). Swin-unet: unet-like pure transformer for medical image segmentation. *arXiv [Preprint] arXiv*:2105.05537. doi: 10.48550/arXiv.2105.05537

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv [Preprint] arXiv*:2102.04306. doi: 10.48550/arXiv.2102.04306

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous

convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848. doi: 10.1109/TPAMI.2017.2699184

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 801–818. doi: 10.1145/3065386

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Istanbul: Springer), 424–432.

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., et al. (2018). "Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (Washington, DC: IEEE), 168–172.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., et al. (2017). "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 764–773.

Dai, Y., Gao, Y., and Liu, F. (2021). Transmed: Transformers advance multi-modal medical image classification. *Diagnostics* 11, 1384. doi: 10.3390/diagnostics11081384

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL: IEEE), 248–255

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv [Preprint] arXiv*:2010.11929. doi: 10.48550/arXiv.2010.11929

Fu, H., Cheng, J., Xu, Y., Wong, D. W. K., Liu, J., and Cao, X. (2018). Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE Trans. Med. Imaging* 37, 1597–1605. doi: 10.1109/TMI.2018.2791488

Fu, S., Lu, Y., Wang, Y., Zhou, Y., Shen, W., Fishman, E., et al. (2020). "Domain adaptive relational reasoning for 3d multi-organ segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Lima: Springer), 656–666.

Gehlot, S., and Gupta, A. (2021). "Self-supervision based dual-transformation learning for stain normalization, classification andsegmentation," in *International Workshop on Machine Learning in Medical Imaging* (Strasbourg: Springer), 477–486.

Gehlot, S., Gupta, A., and Gupta, R. (2020). "Ednfc-net: Convolutional neural network with nested feature concatenation for nuclei-instance segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 1389–1393.

Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., et al. (2019). Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Trans. Med. Imaging* 38, 2281–2292. doi: 10.1109/TMI.2019.2903562

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., et al. (2019). "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul: IEEE), 1314–1324.

Hu, H., Gu, J., Zhang, Z., Dai, J., and Wei, Y. (2018). "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 3588–3597.

Hu, H., Zhang, Z., Xie, Z., and Lin, S. (2019). "Local relation networks for image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (Seoul) 3464–3473.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 4700–4708.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., et al. (2020). "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)* (Barcelona: IEEE), 1055–1059.

Ji, J., Lu, X., Luo, M., Yin, M., Miao, Q., and Liu, X. (2020). Parallel fully convolutional network for semantic segmentation. *IEEE Access* 9, 673–682. doi: 10.1109/ACCESS.2020.3042254

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105.

Kumar, P., Nagar, P., Arora, C., and Gupta, A. (2018). "U-segnet: fully convolutional neural network based automated brain tissue segmentation tool," in *2018 25th IEEE International Conference on Image Processing (ICIP)* (Athens: IEEE), 3503–3507.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., and Heng, P.-A. (2018). H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Trans. Med. Imaging* 37, 2663–2674. doi: 10.1109/TMI.2018.2845918

Liang, D., Qiu, J., Wang, L., Yin, X., Xing, J., Yang, Z., et al. (2020). Coronary angiography video segmentation method for assisting cardiovascular disease interventional treatment. *BMC Med. Imaging* 20, 1–8. doi: 10.1186/s12880-020-00460-9

Liang, J., Yang, C., Zeng, M., and Wang, X. (2022). Transconver: transformer and convolution parallel network for developing automatic brain tumor segmentation in mri images. *Quant. Imaging Med. Surg.* 12, 2397. doi: 10.21037/qims-21-919

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 2117–2125.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv [Preprint] arXiv*:2103.14030. doi: 10.1109/ICCV48922.2021.00986

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford, CA: IEEE), 565–571.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention u-net: learning where to look for the pancreas. *arXiv [Preprint] arXiv*:1804.03999. doi: 10.48550/arXiv.1804.03999

Patil, D. D., and Deore, S. G. (2013). Medical image segmentation: a review. *Int. J. Comput. Sci. Mobile Comput.* 2, 22–27.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., and Jagersand, M. (2020). U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognit.* 106, 107404. doi: 10.1016/j.patcog.2020.107404

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 4510–4520.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., et al. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* 53, 197–207. doi: 10.1016/j.media.2019.01.012

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv [Preprint] arXiv:*1409.1556. doi: 10.48550/arXiv.1409.1556

Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 5693–5703.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV,: IEEE), 2818–2826.

Tan, M., and Le, Q. (2019). "Efficientnet: rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning* (Taiyuan: PMLR), 6105–6114.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). "Training data-efficient image transformers distillation through attention," in *International Conference on Machine Learning* (PMLR), 10347–10357.

Vasuda, P., and Satheesh, S. (1713). Improved fuzzy c-means algorithm for mr brain image segmentation. *Int. J. Comput. Sci. Eng.* 2, 2010.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, (Long Beach, CA), 5998–6008.

Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.-H., Chen, Y.-W., et al. (2021a). Mixed transformer u-net for medical image segmentation. *arXiv [Preprint] arXiv:*2111.04734. doi: 10.1109/ICASSP43922.2022.9746172

Wang, R., Cao, S., Ma, K., Zheng, Y., and Meng, D. (2021b). Pairwise learning for medical image segmentation. *Med. Image Anal.* 67, 101876. doi: 10.1016/j.media.2020.101876

Wang, R., Zhou, Y., Zhao, C., and Wu, H. (2015). A hybrid flower pollination algorithm based modified randomized location for multi-threshold medical image segmentation. *Biomed. Mater. Eng.* 26, S1345-S1351. doi: 10.3233/BME-151432

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 7794–7803.

Xiao, X., Lian, S., Luo, Z., and Li, S. (2018). "Weighted res-unet for high-quality retina vessel segmentation," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)* (Hangzhou,: IEEE), 327–331.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). "Segformer: simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems*. 34. doi: 10.48550/arXiv.2105.15203

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 1492–1500.

Xu, X., Feng, Z., Cao, C., Li, M., Wu, J., Wu, Z., et al. (2021). An improved swin transformer-based model for remote sensing object detection and instance segmentation. *Remote Sens.* 13, 4779. doi: 10.3390/rs132 34779

Zhang, Y., Liu, H., and Hu, Q. (2021). "Transfuse: fusing transformers and cnns for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Strasbourg: Springer), 14–24.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). "Unet++: a nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Granada: Springer), 3–11.

Zhu, X., Hu, H., Lin, S., and Dai, J. (2019). "Deformable convnets v2: more deformable, better results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 9308–9316.

Conflict of Interest: QG and TT were employed by Imperial Vision Technology.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.