# An Efficient Honeycomb Lung Segmentation Network Combining Multi-Paradigms Representation and Cascade Attention

Bingqian Yang, Xiufang Feng, Yunyun Dong*
School of Software, Taiyuan University of Technology, Taiyuan Shanxi 030024, China

***Abstract*—Honeycomb lung is a pulmonary manifestation that occurs in the terminal stage of various lung diseases, which greatly threatens patients. Due to the different locations and irregular shapes of lesions, the accurate segmentation of the honeycomb region is an essential and challenging problem. However, most deep learning methods struggle to effectively utilize both global and local information from lesion images, resulting in cannot to accurately segment the lesion. In addition, these methods often ignore some semantic information that is necessary for the segmentation of lesion location and shape in the decoding stage. To alleviate these challenges, in this paper, we propose a dual-branch encoder and cascaded decoder network (DECDNet) for segmenting honeycombs lesions. First, we design a dual-branch encoder consisting of ResNet34 and Swin-Transformer with different paradigm representations to extract local features and long-range dependencies respectively. Next, to further combine the different paradigm features, we develop the feature fusion module to obtain richer representation information. Finally, considering the problem of information loss during the decoder, a cascaded attention decoder is constructed to aggregate the multi-stage encoder information to get the final segmentation result. Experimental results demonstrate that our method outperforms other methods on the in-house honeycomb lung dataset. Notably, compared with the other nine universal methods, the proposed DECDNet obtains the highest IoU (86.34%), Dice (92.66%), Precision (93.21%), Recall (92.13%), F1-Score (92.66%), and achieves the lowest HD95 (7.33) and ASD (2.30). In particular, our method enables precisely segmenting lesions under different clinical scenarios as well. Our code and dataset are available at https://github.com/ybq17/DECDNet.**

***Keywords*—Honeycomb lung; attention; convolutional neural network; transformer; image segmentation**

## I. INTRODUCTION

Honeycomb lung, also called interstitial pulmonary fibrosis, is a disease where the lung tissue is destroyed and fibrosed. It exhibits distinctive honeycomb-like features, that seriously threaten patient's life [1-2]. Based on published literature, the annual incidence of honeycomb lung is estimated to be between 0.9 and 13 cases per 100,000 individuals [3]. The survival time from initial diagnosis to death is short, ranging from three to five years, with a poor prognosis and high mortality rate [4-5]. Early diagnosis is vital for improving patient prognosis and prolonging their survival [6]. With the development of radiological technology, computed tomography (CT) has become the gold standard for diagnosing honeycomb lung [7]. Information such as the size and location of lesions in CT images can help doctors identify honeycomb regions and make subsequent treatment plans.

In clinical practice, the contours of lesions are usually delineated by physicians. However, the unique characteristics of honeycomb lung make the delineation of lesions a challenging task for physicians. Honeycomb lesions usually have the following properties: (1) the location of the lesion is not fixed and the contour is complex. (2) The lesions are blended with surrounding normal tissue, resulting in blurred boundaries. (3) The size and shape of lesions often vary from person to person. The above properties make it very time-consuming for doctors to contour lesions manually. In addition, due to the differences in doctor experience, the quality of delineation of diseased regions is inequality. To alleviate the burden on doctors and aid in diagnosis, many studies pay attention to computer-aided automatic diagnosis of lesions [8-12]. However, due to the inherent properties of these methods, it is challenging to achieve an accurate diagnosis of the lesion tissue.

With the development of deep learning, convolutional neural networks (CNN) have achieved significant success in the field of medical imaging due to their powerful feature extraction capabilities [13]. Many CNN methods have been applied to medical image segmentation, such as U-Net [14], V-Net [15], and ResUnet [16] which have obtained excellent results. The above methods all adopt an encoder-decoder structure, where the encoder is used to extract feature information at different scales of the image, and the decoder restores the image according to the encoder information. Since the superiority of this architecture, a few variants based on it have a dominant position in segmentation tasks [17-18]. Despite these models having achieved significant performance, their performance is still restricted due to their inherent receptive field [19]. To overcome the limitations, some works try to integrate attention mechanisms and expand the receptive field to improve segmentation accuracy [20-22]. Although the above works improve the segmentation performance of the network, they still suffer from capturing insufficient long-range dependencies.

Recently, transformer has achieved promising results in the field of natural language processing [23]. Due to its ability to capture long-distance dependencies, it has attracted the attention of researchers, and more and more works attempts to apply it to the field of computer vision. Dosovioskiy [24] was the first to introduce the transformer into the image recognition task; he divided an image into non-overlapping patches instead of tokens and fed them into transformer. In order to further improve the performance of image tasks, multi-scale transformer has emerged. For instance, Swin-Transformer [25] and PVT [26] used sliding windows and pyramid architectures respectively to reduce computational cost. Inspired by these studies, we apply the transformer to the honeycomb lung image segmentation task, but the results are not satisfactory due to the limitations of only using self-attention, which restricts the acquisition of local information.

CNN and transformer focus on two aspects of image information. On the one hand, due to the use of convolutional operations with inductive bias, CNN has locality and translation invariance [27]. This property allows CNN to preferably extract local information, but it also limits the receptive field, resulting in cannot to extract global features. Many solutions have been proposed to solve this problem, such as dilated convolution [28], large kernel convolution [29], and pyramid pooling [30]. However, these approaches can only alleviate this problem and cannot completely solve it. On the other hand, transformer utilizes the self-attention mechanism to capture long-range dependencies, but it neglects local information, resulting in the loss of detailed features.

According to the above analysis, we believe that CNN and transformer can be combined to compensate for their weaknesses and obtain higher performance. Several methods have tried to combine CNN and transformer to get local information and long-range dependencies to segment specific medical images, such as TransUNet [31], Tfcns [32], and HiFormer [33]. However, these architectures have some drawbacks that limit their ability to achieve better performance: 1) They cannot effectively combine local and global information from CNN and transformer respectively. 2) They cannot properly aggregate multi-stage information during the decoding stage. Considering these problems, we propose a novel network named DECDNet that combines different paradigms representation to segment the honeycomb region. Concretely, we first design two encoder branches, one is CNN to extract local information, and the other is transformer to get long-range dependencies. Second, we develop a feature fusion module to efficiently combine features from different branches. Lastly, to better aggregate multi-stage information, we construct an attention cascade decoder called ACD. Our contribution can be summarized as five-fold:

- We innovatively segment large-scale honeycomb lung CT images and plan to open this dataset. Our dataset will be available at https://github.com/ybq17/DECDNet.
- In order to extract the local information and global information of the image, a dual-branch encoder composed of ResNet34 and Swin-Transformer is designed to obtain the multi-paradigms representation of the image.
- Considering the problem of insufficient feature fusion, we develop a feature fusion module for efficiently combining information from different branches.
- To track the information loss in the decoding stage, a novel attention-based cascade decoder is constructed to aggregate multi-stage encoder information.
- Experimental results demonstrate that our proposed DECDNet surpasses other segmentation models as well as adaptable to different clinical scenarios.
- The rest of this paper is organized as follows. We first review related work in Section II and describe the overall architecture in Section III. In Section IV, we evaluate the performance of DECDNet on the honeycomb lung dataset. In Section V we discuss the experimental results. We conclude the paper in Section VI.

## II. Related Works

### A. Honeycomb CT Image Segmentation

CT imaging is a common technology used in clinical practice to diagnose honeycomb lung [34]. Different from pulmonary function tests (PFTs) and composite physiologic index (CPI) assessment in judging the patient's condition [35-36], CT imaging quantifies the degree of fibrosis in the lung and helps physicians diagnose the disease more intuitively [37]. Currently, many related studies focus on automatically segmenting honeycomb regions in CT images to quantify the degree of lesions.

For automatic segmentation methods of honeycomb lung can be divided into the following two categories: traditional computer-based CT analysis and deep learning-based models. The first category uses pulmonary fibrosis disease programs such as CALIPER to perform quantitative analysis of lesions. For instance, Jacob et al., [38] extracted characteristic manifestations in CT images, such as honeycomb lesions, reticular patterns, and ground-glass opacity. Then they used an automated procedure to quantify the severity. Nakagawa et al., [39] deployed a computer-assisted method to quantify fibrosis based on the estimated area of the lesions. However, such methods heavily rely on subjective experience that may lead to suboptimal results. Compared with traditional auxiliary approaches, deep learning-based methods achieve end-to-end automatic segmentation, not only reducing dependence on feature selection but also achieving promising results. Handa et al., [40] developed new deep-learning software that can automatically identify and quantify subdivided parenchymal honeycomb lesions. Su et al., [41] proposed a network called RDNet that accurately segmented the honeycomb regions, and the corresponding Dice index was 0.747. Considering more noise and lower contrast honeycomb images, Wei et al.,[42] introduced the popular transformer to quantify the lesions and achieved remarkable performance. Motivated by these studies, our work emphasizes the use of deep learning-based methods to reduce the role of feature selection and perform end-to-end lesions segmentation.

### B. Convolutional Neural Network

With the development of deep learning, convolutional neural networks (CNN) have achieved remarkable success in a variety of computer vision tasks. In the segmentation task, Long et al., [43] proposed a fully convolutional neural network that achieves pixel-level classification. Inspired by the success of FCN, several approaches were introduced to improve segmentation performance, such as dilated convolution [28] and context modeling [30]. Later, an encoder-decoder network called UNet [14] was proposed by Ronneberger for medical image segmentation. With the popularity of UNet, a series of U-shaped architectures have been developed to better segment medical images, such as UNet++ [44], ResUnet [16], etc. In order to further improve segmentation performance, attention mechanisms were introduced into UNet by Ozan [21]. Attention mechanisms can selectively balance the importance of different spatial locations in the feature maps, allowing the network to focus on relevant objective regions. This approach has shown great potential in improving segmentation performance.

### C. Vision Transformer

Transformer was designed originally for Natural Language Processing (NLP) and has achieved remarkable progress in the field [23]. Instead of using convolutional operations, transformer uses self-attention to capture long-range dependencies. Inspired by the success of transformer in the NLP field, Dosovitskiy et al., [24] proposed the vision transformer (ViT) model, which applied transformer to visual tasks for the first time. ViT split the image into patches, similar to words, and fed them into the transformer. Experimental results demonstrated that ViT surpasses many CNN models and has become the backbone for vision tasks. However, ViT requires a large amount of data and high computational complexity. To address this issue, several methods try to reduce the computational cost of ViT. Wang et al., [25] proposed the Pyramid Vision Transformer (PVT), which is inspired by the CNN pyramid structure and reduced the computational cost by using a hierarchical feature extraction strategy. Liu et al., [26] proposed the Swin Transformer, which only focuses on each local window to reduce the computational cost to linear. Recently, massive work has applied transformer to medical image segmentation tasks. Chen et al., [31] proposed TransUnet, which was the first to apply transformer to medical image segmentation. It used a multi-layer CNN-transformer encoder to extract both local and global information and a CNN decoder to restore the size of the output. Swin-UNet [45] and DS-TransUnet [46] used a pure transformer encoder-decoder architecture for 2D segmentation but did not achieve significant performance improvements. UNETR [47] adopted a transformer encoder to extract information and a CNN decoder to obtain 3D segmentation results.

## III. Methodology

In this section, we describe our proposed method. Firstly, we illustrate a dual-branch encoder composed of ResNet34 and Swin-Transformer to extract complementary local information and long-distance dependencies of the image respectively. In the ResNet34 branch, the texture and structural information of lesions are fully utilized. In the Swin-Transformer branch, more attention is paid to the global features of lesions in images such as location and size. Then, we describe a feature fusion module that can fuse information from different branches at different scales. Finally, we present an ACD decoder that alleviates information loss during the decoding stage. As shown in Fig. 1, the input CT image size is 224×224×3, and the output result is a binary lesion segmentation result. To learn different paradigms representation, the input image is extracted by ResNet34 and multi-level Swin-Transformer respectively to extract multi-scale features of the image. Then, multi-scale features from different branches are fused by four feature fusion modules to obtain richer information. Lastly, to further reduce the information loss during the decoding stage, the ACD decoder receives fusion features to get the final segmentation result. More details are described in the following subsection.

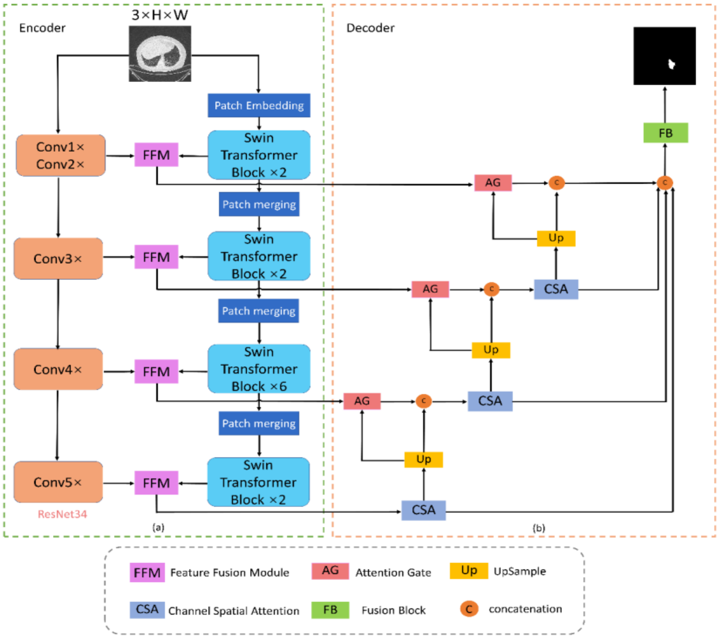

Fig. 1. The architecture of the proposed honeycomb lung segmentation network (DECDNet).

### A. Encoder

Our proposed encoder consists of two multi-scale branches Resnet34, Swin-Transformer, and four feature fusion modules (FFM). Specifically, ResNet34 and Swin-Transformer respectively extract local and global information of images at different scales, such as texture, position, and structure. Here, ResNet34 and Swin-Transformer have four distinct layers each. To produce features with different scales, the patch merging operation will be performed before the feature map is inputted into the next Swin-Transformer layer. Then, the obtained multi-scale features through the above two parallel hierarchical branches contain different levels of semantic information. The FFM fuses each layer's information from different branches to enrich the features and feed them to the decoder.

*1) CNN branch:* As shown in Fig. 1(a), we use ResNet34 to construct the first encoder branch to obtain spatial detail and contextual information in the honeycomb lung images. In this branch, ResNet34 is divided into five blocks, noted as conv1, conv2, conv3, conv4, and conv5. When a feature map goes through a block, its width and height are reduced by a factor of 2. Taking a honeycomb lung CT image of size H×W×3 goes through conv1 and conv2 layers, which will yield a 3D feature $c_1$ of size H/4×W/4×C, where we set C to 48 to ensure feature consistency. Next, by passing through conv3, conv4, and conv5 layers, three features $c_2$, $c_3$, and $c_4$ are obtained, with sizes of H/8×W/8×2C, H/16×W/16×4C, and H/32×W/32×8C, respectively. These feature maps contain multi-scale semantic information and will be used as inputs for the feature fusion module.

*2) Transformer branch:* Recently, Alexely [24] proposed ViT, the first application of the transformer model to visual tasks. ViT has achieved outstanding progress in visual tasks due to its ability to capture long-range dependencies. It mainly consists of two parts: multi-head self-attention (MSA) and multi-layer perception (MLP). However, it has a major issue that its exponential computational complexity makes it difficult for downstream tasks such as image segmentation. To address this problem, Swin-Transformer [25] is developed that only focuses on local regions, greatly reducing the computational complexity. Inspired by the success of Swin-Transformer, we stack 12 Swin-Transformer blocks and patch operations to build the second encoder branch. In this branch, the focus is more on capturing the location and size information of the honeycomb lesions. Each Swin-Transformer is composed of two modules. The first module consists of W-MSA, LN, MLP, and residual connections. The second module introduces a sliding window mechanism to improve W-MSA and enhance information interaction. This process is defined as follows.

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

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

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

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

The encoder branch is divided into four layers, as shown in Fig. 1(a). The first layer consists of two Swin-Transformer blocks and a patch embedding operation that splits the image into patches. The remaining three layers use patch merging to reduce the size of the image. And the number of Swin-Transformer blocks is set to 2, 6, and 2 for these layers, respectively. In this branch, the input image has the same dimensions as the CNN branch, which is H×W×3. The outputs of each layer are denoted as $t_1$, $t_2$, $t_3$, and $t_4$, with sizes of ((H/4, W/4), C), ((H/8, W/8), 2C), ((H/16, W/16), 4C) and ((H/32, W/32), 8C) respectively. They have the same pixel points as the outputs of the corresponding level CNN branch. We set the patch size to 4, so the feature dimension C is equal to 4×4×3=48.

*3) Feature fusion module:* In order to fuse different branches of features, we design the feature fusion module called FFM. It can efficiently and flexibly fuse features from CNN and transformer branches with different resolutions and different channels, and its structure is illustrated in Fig. 2. In this module, we first adjust the shape of the transformer features by reshaping operation. Then the CNN features and transformer features (called f and g) from each encoder branch are adjusted for dimensions by 1*1 convolution. Next, the f and g are merged by concatenation operation. The merged features are fed into the 1×1 convolution, and then the normalization operation and activation function are executed. Finally, the two encoder branch features are completely fused by a 1×1 convolution layer. In the encoder, we deploy four feature fusion modules to fuse the multi-scale features of CNN and Transformer under different branches. The fused feature balances multi-scale global information and local features, enriching the representation information.

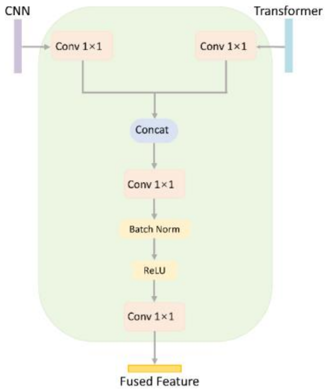

Fig. 2. The structure of the proposed Feature-Fusion Module (FFM).

### B. Decoder

As illustrated in Fig. 1(b), we propose a novel decoder called ACD that can efficiently aggregate multi-stage information from the encoder. It consists of four components: Up, AG, CSA, and FB. Up is used for upsample operation, AG is used for cascaded feature fusion, CSA can refine features, and FB is used to fuse multi-level features. Specifically, we set up three CSA to enhance feature information at different scales and three AG corresponding to the output of FFB. In addition, AG is employed to integrate the fusion information from the FFB and the upsampled features from the lower level. Next, we use a concatenation operation to merge the features of AG and the lower layer. Then, CSA is executed to refine the merged features. Finally, we use FB to integrate the output of different layers to obtain the final prediction. More details are described in the following subsection.

*1) Up:* Up is used to restore the current feature size to match the dimension of the upper layer's feature. Each Up consists of a ReLU activation function, batch normalization operation, a 3*3 convolution operation, and a linear upsampling. Upsample () with an upsampling factor of 2. The Up operation can be formulated as follows:

$$Up(x) = (ReLu(BN(Conv(Upsample(x))))) \qquad (2)$$

*2) Fusion block:* To enhance feature representation, we design the Fusion Block (FB) to efficiently utilize multi-level features. As shown in Fig. 3(a), FB consists of two parts: residual connection and Convblock. The Convblock consists of Convolution (Conv), Batch Normalization (BN), and Rectified Linear Unit (ReLu). On the one hand, the convolutional block assists the network in enhancing features. On the other hand, the residual connection avoids gradient explosion. Then, the results of the two parts are added together to obtain the final prediction result.

*3) Channel spatial attention:* The Channel Spatial Attention (CSA) module can refine the feature map, as shown in Fig. 3(b). It consists of spatial attention, channel attention, and a 1×1 convolution. When the input features enter this module, parallel channel attention and spatial attention are used to enhance the spatial and channel features of the image, respectively. Then, these features are fused through concatenation and the dimension is reduced through convolution. The fusion feature retains important spatial and channel information, which assists the decoder in better-recovering information.

*4) Attention gate:* Motivated by the success of the attention mechanism, we introduce the attention gate (AG), which can combine multi-stage features, extract areas of interest, and ignore irrelevant parts using spatial attention. As shown in Fig. 3(c), each AG consists of two 1x1 convolutions to change dimensions, two batch normalizations, and two activation functions: ReLu and Sigmoid. Specifically, the FFB features denoted f is first added point-wise with the features from the lower-level denoted d. Then, convolution and activation operations are performed to obtain the spatial attention map. Finally, the attention map is element-wise multiplied by f using the Hadamard product.

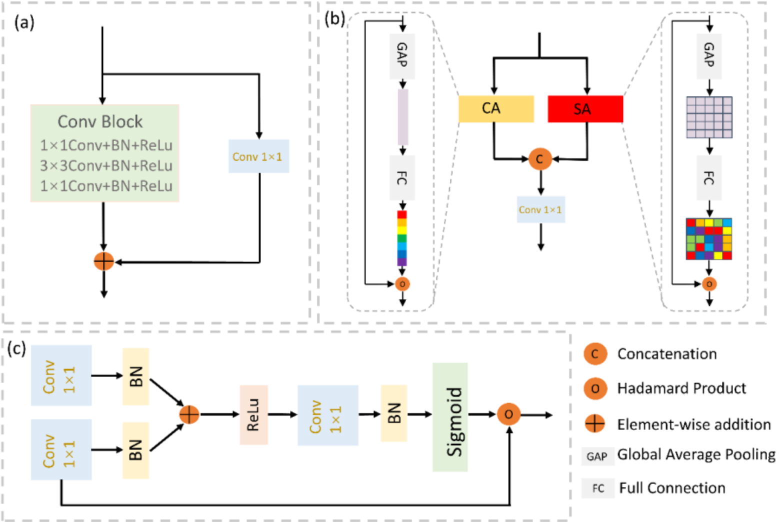

Fig. 3. Structure of (a) FB, (b) CSA, and (c) AG.

## IV. Experimental Results

### A. Datasets

We use CT images of honeycomb lungs provided by Shanxi Provincial People's Hospital for the experiment. A total of 7170 honeycomb images with size 512 × 512 pixels were collected from chest CT scans of 121 patients as our dataset. These scans were conducted using specific scan parameters, including an X-ray voltage of approximately 120 kVp, a current of 200-500 mA, and a gantry rotation speed of 0.5 seconds per rotation. And the slice images were acquired with a uniform slice thickness of 5 mm. The dataset includes images with lesion sizes ranging from small focal areas of a few millimeters to extensive lesions spanning large lung areas. Lesion appearances also vary, from early-stage subtle honeycombing with fine reticular patterns to advanced-stage prominent cystic changes as illustrated in Fig. 4. All of these images are labeled by experienced radiologists. After annotation, we resize the images to 224×224 and apply normalization to accelerate model convergence. We divide the dataset into training/validation/testing sets at a ratio of 6:2:2, with 4302 images used for training and 1434 images used for validation and testing.

### B. Evaluation Metrics

We adopt five common metrics to evaluate segmentation performance: Dice coefficient (Dice), Jaccard index (IoU), Precision, Recall, and F1-score. These metrics are calculated using true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN):

$$Dice = \frac{2TP}{2TP+FP+FN} \tag{3}$$

$$IoU = \frac{TP}{TP+FP+FN} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

$$F1-score = \frac{2\times Recall\times Precision}{Recall+Precision} \tag{7}$$

In addition, Hausdorff distance (HD) 95% and average surface distance (ASD) are used to measure the performance of the model in segmenting the boundaries of honeycomb regions.

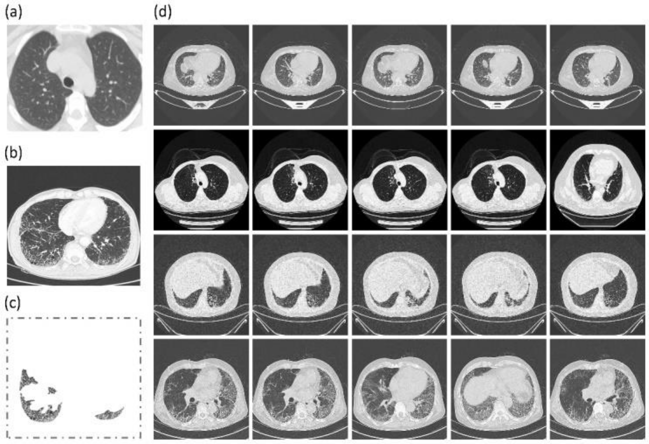

Fig. 4. (a) A normal lung image, (b) A honeycomb lung image, (c) Honeycomb lung lesions, and (d) Some honeycomb lung images in the datasets.

### C. Implementation Details

Our architecture is implemented using PyTorch and trained on an RTX Nvidia 3090 GPU. And we use the SGD as the optimizer with a momentum of 0.9 and weight decay of 0.0001. The learning rate and batch size are set to 0.0001 and 24, respectively. Further, aiming to improve the generalization and robustness of the model, we use data augmentation technology such as flipping and rotation to diversify the data.

The loss function is composed of commonly used cross-entropy loss and dice loss, defined as follows:

$$= \alpha l_1 + (1-\alpha) l_2 \tag{8}$$

where, $l_1$ denotes the cross-entropy loss and $l_2$ denotes the dice loss. The weight ratio α is a balancing factor, set to 0.4 based on experimental testing.

### D. Evaluation Results

To demonstrate the effectiveness of our proposed DECDNet, we conducted comparative experiments with nine methods. All methods are evaluated both quantitatively and qualitatively. These competing methods include the following: U-Net [12], Att-UNet [19], UNet++ [42], FPN [28], DABNet [46], ViT [22], CGNet [47], TransUNet [29], SwinUNet [43]. The above methods cover three types: traditional CNN architecture, pure transformer architecture, and the combined architecture of CNN and transformer. And we evaluated our method on seven evaluation metrics mentioned earlier. The smaller the ASD and HD95 value is, the better the performance is; the larger the other index values are, the better the performance is. Table I shows the qualitative evaluation results of all methods on the honeycomb lung dataset. The experimental results demonstrate that our proposed network has higher IoU, Dice, Precision, Recall, and F1-score, as well as smaller HD95 and ASD. Compared with the second-ranked SwinUNet, our method increases IoU and Dice by 2.13% and 1.23%, respectively, reaching 86.34% and 92.66% Precision, Recall, and F1-Score also improved by 1.68%, 0.81%, and 1.24%, respectively. HD95 and ASD decreased by 2.08 and 0.51 compared to Swin-UNet, reaching 7.33 and 2.30, respectively. These results indicate that our proposed DECDNet can accurately segment the honeycomb lesions. Moreover, our method outperforms traditional CNN methods, ViT, and TransUnet in terms of FLOPs or Params, not only achieving the best segmentation results but also balancing the accuracy and computational complexity.

Meanwhile, to quantify segmentation performance, we conducted visual comparisons of the segmentation with different methods on the honeycomb lung dataset. As shown in Fig. 5, we visually compared the segmentation results of UNet, FPN, TransUNet, SwinUNet, and our model with the ground truth. Within these results, the red box draws attention to significant differences compared with the mask. Obviously, the transformer method is more accurate than the traditional CNN method for segmenting honeycomb lung lesions However, due to the transformer ignores the local features of the image, the segmentation of the lesion boundary is smoother compared with the mask. In contrast, our proposed architecture efficiently utilizes different paradigms representation and alleviates the information loss during the decoding stage. The boundary delineation and region segmentation of the honeycomb lung are more similar to the ground truth. In the above quantitative and qualitative analyses, DECDNet achieves the best performance on the honeycomb lung dataset, demonstrating that our proposed method can accurately segment the contours and regions of honeycomb lung.

Additionally, to evaluate the adaptability of this model in clinical practice, we performed experiments to evaluate our method under various conditions considering the patient positioning, lesion size, and the presence of adjacent organs that commonly occur in clinical applications. We conducted the visual analysis of CT images of different lesion sizes, different slices, and different angles. As shown in Fig. 6 and Fig. 7, regardless of the size, axis, or slice of the lesion, DECDNet can accurately segment and outline the lesions. Table II shows the quantitative results for Dice and Hd95 in various conditions, further proving the adaptability of our method for lesion area segmentation in different clinical scenarios.

TABLE I. COMPARING OUR METHOD WITH OTHER METHODS ON THE HONEYCOMB LUNG DATASET

| Methods | IoU (%) | Dice (%) | Precision (%) | Recall (%) | F1-Score (%) | HD | ASD | FLOPs(G) | Params(M) |
|---|---|---|---|---|---|---|---|---|---|
| U-Net [14] | 76.82 | 86.89 | 87.58 | 86.20 | 86.38 | 15.05 | 5.27 | 37.03 | 31.04 |
| Att-UNet [21] | 78.12 | 87.72 | 87.21 | 88.23 | 87.71 | 14.20 | 4.54 | 64.19 | 57.16 |
| U-Net++ [44] | 79.67 | 88.68 | 87.71 | 89.68 | 88.68 | 13.69 | 4.13 | 103.36 | 47.19 |
| FPN [30] | 81.65 | 89.90 | 89.89 | 89.46 | 89.67 | 12.11 | 3.86 | 38.49 | 36.77 |
| DABNet [48] | 82.08 | 90.16 | 90.90 | 89.32 | 90.10 | 11.58 | 3.52 | 1.01 | 0.75 |
| ViT [24] | 80.32 | 89.08 | 89.48 | 88.57 | 89.02 | 13.15 | 4.78 | 17.58 | 85.80 |
| CGNet [49] | 82.72 | 90.54 | 91.01 | 89.38 | 90.18 | 11.27 | 3.39 | 0.68 | 0.49 |
| TransUNet [31] | 83.53 | 91.03 | 91.55 | 90.49 | 91.02 | 10.36 | 2.97 | 29.35 | 105.28 |
| SwinUNet [45] | 84.21 | 91.43 | 91.53 | 91.32 | 91.42 | 9.41 | 2.81 | 6.16 | 27.17 |
| Ours | 86.34 | 92.66 | 93.21 | 92.13 | 92.66 | 7.33 | 2.30 | 17.62 | 90.37 |

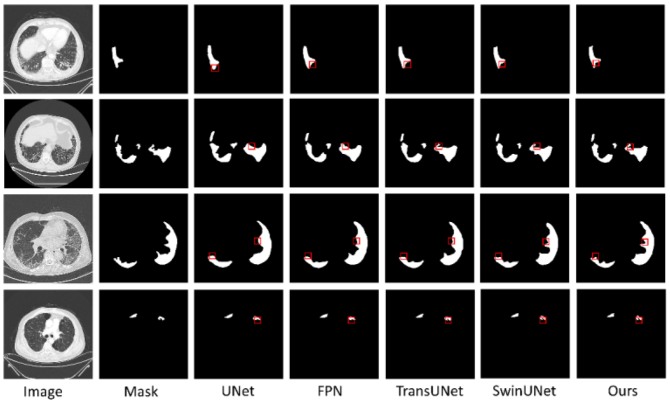

Fig. 5. The qualitative comparison of the honeycomb lung dataset.

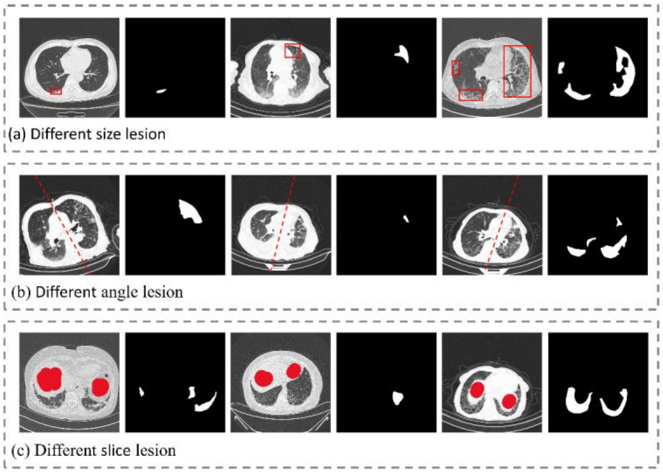
(a) Different size lesion
(b) Different angle lesion
(c) Different slice lesion

Fig. 6. DECDNet segmentation lesions visualization under different conditions (a) represents the lesion size (b) represents scan angle (c) represents the slice lesion.

TABLE II. QUANTITATIVE ANALYSIS UNDER DIFFERENT CONDITIONS

| Condition | Dice | HD |
|---|---|---|
| Different size | 91.24 | 7.61 |
| Different angle | 92.83 | 7.15 |
| Different slice | 92.70 | 7.26 |

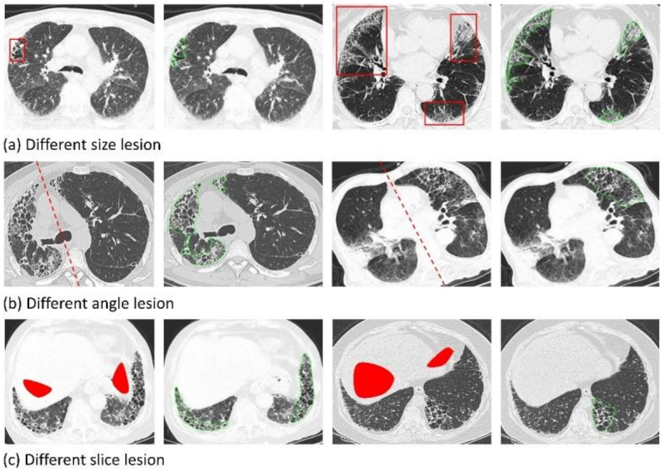
(a) Different size lesion
(b) Different angle lesion
(c) Different slice lesion

Fig. 7. DECDNet segmentation lesion boundaries visualization under different conditions (a) represents the lesion size (b) represents scan angle (c) represents the slice lesion.

## *E. Ablation Studies*

*1) Evaluation of encoder:* Our encoder is composed of CNN and transformer with different paradigms. Since the transformer can capture long-range dependencies, it tends to ignore local feature information. To address this issue, we introduced CNN as an auxiliary branch to compensate for the loss of spatial detail information. To further verify the importance of the CNN branch, we employed variants of pure transformer and CNN-combined transformer as two encoder structures. As shown in Fig. 8, the encoder that combines CNN with the transformer outperforms the pure transformer encoder. The results presented in Table III suggest that the integration of the CNN branch also enhances boundary segmentation performance. Thus, the introduction of CNN as an auxiliary branch is necessary to improve segmentation performance.

TABLE III. ABLATION EXPERIMENTS ON BOUNDARY EVALUATION FOR CNN BRANCH

| Methods | HD | ASD |
|---|---|---|
| No-CNN | 8.40 | 3.01 |
| Ours | 7.33 | 2.30 |

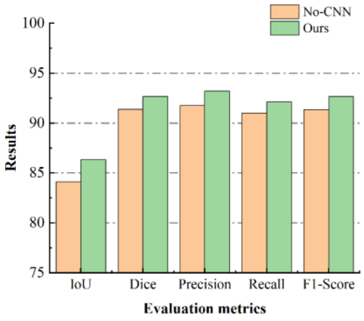

Fig. 8. Ablation study on the influence of CNN branch.

*2) Evaluation of FFM:* To verify the effectiveness of FFM in segmentation performance, we removed FFM and used Conv2d to adjust the feature dimensions, then adopted the add operation to fuse feature. In this way, we denote the variable as Conv2d+add to verify the importance of FFM. The segmentation performance shown in Fig. 9 reveals the effect of FFM in feature fusion. Specifically, the model with FFM achieved higher IoU, Dice, Precision, Recall, and F1-score compared to using only the add operation. In addition, the employment of FFM can improve boundary segmentation performance. As shown in Table IV, our method achieved lower HD95 and ASD. These results demonstrate that simply using add cannot fully utilize the features from CNN and transformer, while FFM can better help the network utilize different branches feature.

*3) Evaluation of decoder:* The decoder is an important factor that affects segmentation performance, and its main role is to restore feature maps to obtain the final prediction. The classic decoder does not use attention mechanisms but restores image resolution by fusing multi-stage skip connections and upsampling features to achieve the final prediction. While, Atten-UNet uses attention in skip connections during the decoding stage to focus on the lesions, denoted as the attention decoder. Different from the above decoder, we proposed a new attention-based cascaded decoder called ACD to efficiently combine multi-stage features from the encoder, enabling the network to better focus on the honeycomb regions. The results shown in Fig. 10 demonstrate that compared to the classic decoder and attention decoder, the architecture using ACD achieves improvements in IoU, Dice, Precision, Recall, and F1-score. Our decoder also obtains the smallest HD95 and ASD, as shown in Table V. The presented results provide evidence that our decoder can better focus on the lesions and efficiently integrate multi-stage features to get segmentation results.

*4) Evaluation of combination FFM and cascade:* To verify the effectiveness of the FFM and Cascade decoder combination for model performance, we conducted four sets of experiments to explore the influence of each component on honeycomb lung segmentation. Firstly, setting a model that only uses pointwise additive fusion and classical encoder for segmentation as the baseline model. To ensure fairness, all experiments adopt the same environment settings.

TABLE IV. ABLATION EXPERIMENTS ON BOUNDARY EVALUATION FOR FFM

| Methods | HD | ASD |
|---|---|---|
| Conv2d+add | 8.27 | 3.09 |
| Ours | 7.33 | 2.30 |

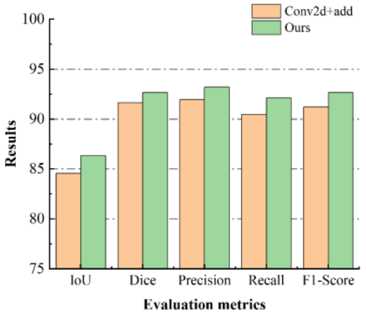

Fig. 9. Ablation study on the influence of FFM.

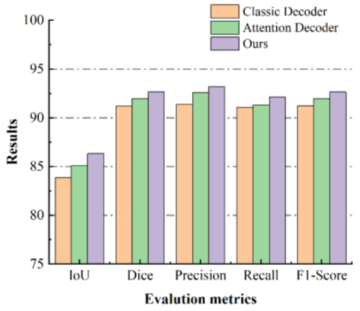

Fig. 10. Ablation study on the influence of different decoders.

TABLE V. ABLATION EXPERIMENTS ON BOUNDARY EVALUATION FOR DIFFERENT DECODERS

| Methods | HD | ASD |
|---|---|---|
| Classic decoder | 11.30 | 3.38 |
| Attention decoder | 8.19 | 2.88 |
| Ours | 7.33 | 2.30 |

As shown in Fig. 11, the combination of FFM and Cascade decoder achieved the best segmentation performance of honeycomb lesions. For boundary outline, the combination method also obtains the lowest HD and ASD as illustrated in Table VI, which is more similar to the groundtruth. Therefore, FFM and Cascade are necessary to improve the accuracy of segmentation.

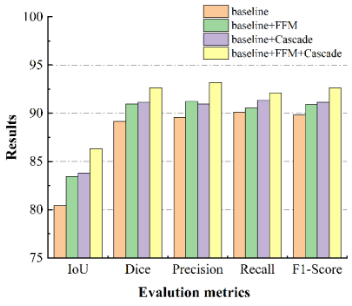

Fig. 11. Ablation study on the influence of the combination FFM and Cascade.

TABLE VI. ABLATION EXPERIMENTS ON BOUNDARY EVALUATION FOR THE COMBINATION FFM AND CASCADE

| Methods | HD | ASD |
|---|---|---|
| baseline | 12.78 | 4.05 |
| baseline+FFM | 10.60 | 3.29 |
| baseline+Cascade | 9.82 | 3.11 |
| baseline+FFM+Cascade | 7.33 | 2.30 |

## V. DISCUSSION

Honeycomb lung is a terminal manifestation of lung disease, which greatly threatens patients. In clinical applications, the segmentation of lesions is essential. It aids in evaluating lesions, identifying the distribution of lesions, and assisting doctors in making accurate diagnoses. Moreover, segmenting honeycomb lung is a challenging task due to their ambiguous and irregular characteristics. Therefore, it is crucial to design a network that achieves higher segmentation accuracy for the precise localization of honeycomb lung lesions. Our proposed DECDNet architecture integrates global and local information from different paradigms to alleviate the limitations of CNN and transformer. In addition, the specially designed ACD decoder can effectively recover image information from the encoder. We conducted experiments on our method and nine universal segmentation algorithms, and our method achieved the highest IoU (86.34%), Dice (91.87%), Recall (92.13%), Precision (93.21%), F1-score (92.66%), and the smallest HD95 (7.33) and ASD (2.30). To quantify the results, we show the visual segmentation results of different methods, as shown in Fig. 5. Compared with other methods, our method can not only focus on the major lesions but also pay more attention to the boundaries of the honeycomb lung. Next, we visualized the segmentation performance in different clinical situations, as shown in Fig. 6 and Fig. 7, indicating that the proposed model is adaptable to diverse clinical scenarios. Additionally, to verify the effectiveness of each part of our architecture, we conducted three groups of ablation experiments to explore the effects of the dual-branch encoder, FFM, and ACD decoder. The results show that all three parts are effective and can improve the segmentation performance of the network. Therefore, our proposed method can precisely segment the honeycomb lung lesions, alleviate the burden on doctors, and assist in diagnosis.

Although our method has shown outstanding performance on the honeycomb lung dataset, it still has some defects that need to be addressed. On one hand, our model did not achieve promising performance in cases where the lesion size is small and the background is complex. On the other hand, as we have only tested on data from a single center, the generalizability of our model remains to be considered. In the future, we will expand and improve our method by adjusting multi-scale inputs and collecting data from multiple centers to address these issues.

## VI. CONCLUSION

In this paper, we propose a novel network called DECDNet for the segmentation of honeycomb lung CT images. Specifically, we first design a dual-branch encoder to efficiently capture global and local information from different paradigms. Next, the feature fusion module is developed to fuse CNN and transformer features. Finally, we develop an attention-based cascade decoder to aggregate multi-stage encoder information. Our method demonstrated its effectiveness in extensive experiments through the effective extraction, fusion, and restoration of local information (such as the texture and structure of lesions) and global information (such as location and size). And our model achieves state-of-the-art performance on the honeycomb lung dataset. In addition, our model also accurately segments lesions under various conditions, making it a valuable method for assisting doctors in locating and tracking lesions, as well as making diagnoses. In the future, we will focus on further automatic diagnosis of honeycomb lung, such as by adopting multi-scale inputs to avoid noise and collecting multi-center data to enhance the model's generalizability.

## ACKNOWLEDGMENT

We thank all the anonymous reviewers for their valuable comments and constructive suggestions. This work is supported by the National Natural Science Foundation of China (No.62306206); the Key Research and Development Program of Shanxi Province (No.202102020101007).

## REFERENCES

[1] M. Hosseini and M. Salvatore, “Is pulmonary fibrosis a precancerous disease?” European Journal of Radiology, p. 110723, 2023.

[2] G Raghu, M Remy-Jardin, L Richeldi, C. C. Thomson, Y. Inou, and T. Johkoh, “Idiopathic pulmonary fibrosis (an update) and progressive pulmonary fibrosis in adults: an official ATS/ERS/JRS/ALAT clinical practice guideline,” American Journal of Respiratory and Critical Care Medicine, vol. 205, no. 9, pp. e18–e47, 2022.

[3] T. M. Maher, E. Bendstrup, L. Dron, J. Langley, G. Smith, and J. M. Khalid, “Global incidence and prevalence of idiopathic pulmonary fibrosis,” Respiratory research, vol. 22, no. 1, pp. 1–10, 2021.

[4] Q. Zheng, I. A. Cox, J. A. Campbell, Q. Xia, P. Otahal, and Bde. Graaff, “Mortality and survival in idiopathic pulmonary fibrosis: a systematic review and meta-analysis,” ERJ Open Research, vol. 8, no. 1, 2022.

[5] J. Ji, S. Zheng, Y. Liu, T. Xie, X. Zhu, and Y. Ni, “Increased expression of OPN contributes to idiopathic pulmonary fibrosis and indicates a poor prognosis,” Journal of Translational Medicine, vol. 21, no. 1, p. 640, 2023.

[6] M. B. Herberts, T. T. Teague, V. Thao, L. R. Sangaralingham, H. J. Henk, and K. T. Hovde, “Idiopathic pulmonary fibrosis in the United States: time to diagnosis and treatment,” BMC Pulmonary Medicine, vol. 23, no. 1, p. 281, 2023.

[7] S. Hobbs, J. H. Chung, J. Leb, K. Kaproth-Joslin, and D. A. Lynch, “Practical imaging interpretation in patients suspected of having idiopathic pulmonary fibrosis: official recommendations from the Radiology Working Group of the Pulmonary Fibrosis Foundation,” Radiology: Cardiothoracic Imaging, vol. 3, no. 1, p. e200279, 2021.

[8] P. Mathur, A. S. Raghuvanshi, A. Kumari, and A. Chandra, “Computer-Aided Diagnosis System for Brain Tumor Classification and Segmentation,” in 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN), 2023, pp. 547–552.

[9] Y. Xia, H. Yun, and Y. Liu, “MFEFNet: Multi-scale feature enhancement and Fusion Network for polyp segmentation,” Computers in Biology and Medicine, vol. 157, p. 106735, 2023

[10] X. He, G. Qi, Z. Zhu, Y. Li, B. Cong, and L. Bai, “Medical image segmentation method based on multi-feature interaction and fusion over cloud computing,” Simulation Modelling Practice and Theory, vol. 126, p. 102769, 2023.

[11] Aamir M, Rahman Z, Dayo Z A, Abor W A, Uddin M, and Khan , “A deep learning approach for brain tumor classification using MRI

images," Computers and Electrical Engineering, vol. 101, p. 108105, 2022.

[12] Aamir M, Rahman Z, Abro W A, Bhatti U A, Dayo Z A, and Ishfaq M, "Brain tumor classification utilizing deep features derived from high-quality regions in MRI images," Biomedical Signal Processing and Control, vol. 85, p. 104988, 2023.

[13] P. Malhotra, S. Gupta, D. Koundal, A. Zaguia, and W. Enbeyle, "Deep neural networks for medical image segmentation," Journal of Healthcare Engineering, vol. 2022, 2022.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 2015, pp. 234–241.

[15] F. Milletari, N. Navab, and 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.

[16] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," in 2018 9th international conference on information technology in medicine and education (ITME), 2018, pp. 327–331.

[17] H. Sahli, A. Ben Slama, and S. Labidi, "U-Net: A valuable encoder-decoder architecture for liver tumors segmentation in CT images," Journal of X-ray science and technology, vol. 30, no. 1, pp. 45–56, 2022.

[18] S. Yalçın and H. Vural, "Brain stroke classification and segmentation using encoder-decoder based deep convolutional neural networks," Computers in Biology and Medicine, vol. 149, p. 105941, 2022.

[19] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, 2021, pp. 171–180.

[20] J. Zhang, W. Pan, B. Wang, Q. Chen, and Y. Cheng, "Multi-scale aggregation networks with flexible receptive fields for melanoma segmentation," Biomedical Signal Processing and Control, vol. 78, p. 103950, 2022.

[21] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, and K. Misawa, "Attention u-net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, 2018.

[22] D. P. Fan, G. P. Ji, T. Zhou, G. Chen, H. Fu, and J. Shen, "Pranet: Parallel reverse attention network for polyp segmentation," in International conference on medical image computing and computer-assisted intervention, 2020, pp. 263–273.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and A. N. Gome, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, and T. Unterthiner, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[25] Z Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, and Z. Zhan, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.

[26] W. Wang, E. Xie, X. Li, D. Fan, K. Song, and D. Liang, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 568–578.

[27] A. Mumuni and F. Mumuni, "CNN architectures for geometric transformation-invariant feature representation in computer vision: a review," SN Computer Science, vol. 2, pp. 1–23, 2021.

[28] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122, 2015.

[29] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters--improve semantic segmentation by global convolutional network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4353–4361.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.

[31] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, and Y. Wang, "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.

[32] Z. Li, D. Li, C. Xu, W. Wang, Q. Hong and Q. L, "Tfcns: A cnn-transformer hybrid network for medical image segmentation," in International Conference on Artificial Neural Networks, 2022, pp. 781–792.

[33] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, and J. Cohen-Adad, "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6202–6212.

[34] P. Li, J. Zhang, W. Yua, and S. Yu, "Idiopathic interstitial pneumonia," In: Radiology of Infectious and Inflammatory Diseases-Volume 3: Heart and Chest, 2023, pp. 309–323.

[35] R. Gupta, J. S. Kim, and R. P. Baughman, "An expert overview of pulmonary fibrosis in sarcoidosis," Expert Review of Respiratory Medicine, vol. 17, no. 2, pp. 119–130, 2023.

[36] Q. Wang, Z. Xie, N. Wan, L. Yang, Z. Jin, and F. Jin, "Potential biomarkers for diagnosis and disease evaluation of idiopathic pulmonary fibrosis," Chinese Medical Journal, vol. 136, no. 11, pp. 1278–1290, 2023.

[37] Y. Kunihiro, T. Matsumoto, T. Murakami, M. Shimokawa, H. Kamei, and N. Tanaka, "A quantitative analysis of long-term follow-up computed tomography of idiopathic pulmonary fibrosis: the correlation with the progression and prognosis," Acta Radiologica, p. 02841851231175252, 2023.

[38] J. Jacob, B. J. Bartholmai, S. Rajagopalan, M. Kokosi, A. Nair, and R. Karwoski, "Mortality prediction in idiopathic pulmonary fibrosis: evaluation of computer-based CT analysis with conventional severity measures," European Respiratory Journal, vol. 49, no. 1, 2017.

[39] H. Nakagaw, Y. Nagatani, M. Takahashi, E. Ogawa, N. VanTho, and Y. Ryujin, "Quantitative CT analysis of honeycombing area in idiopathic pulmonary fibrosis: correlations with pulmonary function tests," European Journal of Radiology, vol. 85, no. 1, pp. 125–130, 2016.

[40] T. Handa, K. Tanizawa, T. Oguma, R. Uozumi, K. Watanabe, and N. Tanab, "Novel artificial intelligence-based technology for chest computed tomography analysis of idiopathic pulmonary fibrosis," Annals of the American Thoracic Society, vol. 19, no. 3, pp. 399–406, 2022.

[41] N. Su, F. Hou, W. Zheng, Z. Wu, and E. Linning, "Computed Tomography–Based Deep Learning Model for Assessing the Severity of Patients With Connective Tissue Disease–Associated Interstitial Lung Disease," Journal of computer assisted tomography, vol. 47, no. 5, pp. 738–745, 2023.

[42] W. Jianjian, G. Li, K. He, P. Li, L. Zhang, and R. Wang, "MCSC-UTNet: Honeycomb lung segmentation algorithm based on Separable Vision Transformer and context feature fusion," In Proceedings of the 2023 2nd Asia Conference on Algorithms, Computing and Machine Learning, 2023, pp. 488–494.

[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[44] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," IEEE transactions on medical imaging, vol. 39, no. 6, pp. 1856–1867, 2019.

[45] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, and Q. Tian , "Swin-unet: Unet-like pure transformer for medical image segmentation," in European conference on computer vision, Springer, 2022, pp. 205–218.

[46] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–15, 2022.

[47] A Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, and B. Landman, “Unetr: Transformers for 3d medical image segmentation,” in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 574–584.

[48] G. Li, I. Yun, J. Kim, and J. Kim, “Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation,” arXiv preprint arXiv:1907.11357, 2019.

[49] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, “Cgnet: A light-weight context guided network for semantic segmentation,” IEEE Transactions on Image Processing, vol. 30, pp. 1169–1179, 2020.