

DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation

Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, *Senior Member, IEEE*, and Guangming Lu, *Member, IEEE*

Abstract—Automatic medical image segmentation has made great progress benefit from the development of deep learning. However, most existing methods are based on convolutional neural networks (CNNs), which fail to build long-range dependencies and global context connections due to the limitation of receptive field in convolution operation. Inspired by the success of Transformer whose self-attention mechanism has the powerful abilities of modeling the long-range contextual information, some researchers have expended considerable efforts in designing the robust variants of Transformer-based U-Net. Moreover, the patch division used in vision transformers usually ignores the pixel-level intrinsic structural features inside each patch. To alleviate these problems, in this paper, we propose a novel deep medical image segmentation framework called Dual Swin Transformer U-Net (DS-TransUNet), which might be the first attempt to concurrently incorporate the advantages of hierarchical Swin Transformer into both encoder and decoder of the standard U-shaped architecture to enhance the semantic segmentation quality of varying medical images. Unlike many prior Transformer-based solutions, the proposed DS-TransUNet first adopts dual-scale encoder subnetworks based on Swin Transformer to extract the coarse and fine-grained feature representations of different semantic scales. As the core component for our DS-TransUNet, a well-designed Transformer Interactive Fusion (TIF) module is proposed to effectively establish global dependencies between features of different scales through the self-attention mechanism, in order to make full use of these obtained multi-scale feature representations. Furthermore, we also introduce the Swin Transformer block into decoder to further explore the long-range contextual information during the up-sampling process. Extensive experiments across four typical tasks for medical image segmentation demonstrate the effectiveness of DS-TransUNet, and show that our approach significantly outperforms the state-of-the-art methods.

Index Terms—Medical image segmentation; Long-range contextual information; Hierarchical Swin Transformer; Dual-scale; Transformer Interactive Fusion module

I. INTRODUCTION

MEDICAL image segmentation is an important yet challenging research problem involving many common tasks in clinical applications, such as polyp segmentation, lesion segmentation, cell segmentation, etc. Moreover, medical image segmentation is a complex and key step in the field of medical image processing and analysis, and plays an important

Ailiang Lin, Bingzhi Chen, Jiayu Xu and Guangming Lu are with the Shenzhen Medical Biometrics Perception and Analysis Engineering Laboratory, Harbin Institute of Technology, Shenzhen 518055, China. (e-mail: tianbaoge24@gmail.com, chenbingzhi@stu.hit.edu.cn, jiayuxu1998@gmail.com, luguangm@hit.edu.cn)

Zheng Zhang is with the Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen 518055, China, and also with Shenzhen Key Laboratory of Visual Object Detection and Recognition, Shenzhen 518055, China.(e-mail: darrenzz219@gmail.com)

role in computer-aided clinical diagnosis system. Its purpose is to segment the parts with special significance in medical images and extract relevant features through semi-automatic or automatic process, so as to provide reliable basis for clinical diagnosis and pathological research, and assist doctors in making more accurate diagnosis.

With the development of deep learning, convolutional neural networks (CNNs) have become dominant in a series of medical image segmentation tasks. Among various CNN variants, the typical encoder-decoder based network U-Net [1] has demonstrated excellent segmentation potential, where encoder extracts features through continuous down-sampling, and then decoder progressively leverage features output from encoder through skip connection for up-sampling, so that the network can obtain features of different granularity for better segmentation. Following the popularity of U-Net, many novel models have been proposed such as UNet++ [2], Res-UNet [3], Attention U-Net [4], DenseUNet [5], R2U-Net [6], KiU-Net [7] and UNet 3+ [8], which are specially designed for medical image segmentation and achieve expressive performance. Although CNNs have made great success in the field of medical image, it is difficult for them to make further breakthroughs. Due to the inherent inductive biases, each convolutional kernel can only focus on a sub-region in the whole image, which makes it lose global context and fail to build long-range dependencies. The stacking of convolution layer and down-sampling helps expand the receptive field and bring better local interaction, but this is a sub-optimal choice because it makes the model more complicated and easier to overfit. There exists some works trying to model long-range dependencies for convolution such as attention mechanism [9] [10] [11]. However, since these methods are not aimed at the field of medical image segmentation, they still have great limitations in global context modeling which means there is great potential for improvement.

Recently, the novel architecture Transformer [12] which was originally designed for sequence-to-sequence modeling in natural language processing (NLP) tasks, has sparked tremendous discussion in computer vision (CV) community. Transformer can revolutionize most NLP tasks such as machine translation, named-entity recognition and question answering, mainly because multi-head self attention (MSA) mechanism can effectively build global connection between the tokens of sequences. The ability of long-range dependencies modeling is also suitable for pixel-based CV tasks. Specially, DETR Transformer (DETR) [13] utilizes a elegant design based on Transformer to build the first fully end-to-end object detection model. Vision Transformer (ViT) [14], the first image

recognition model purely based on Transformer is proposed and achieves comparable performance with other state-of-the-art (SOTA) convolution-based methods. To reduce the computational complexity, a hierarchical Swin Transformer [15] is proposed with Window based MSA (W-MSA) and Shifted Window based MSA (SW-MSA) as illustrated in Fig. 1(b), and surpasses the previous SOTA methods in image classification, dense prediction tasks such as object detection and semantic segmentation. SEgmentation TRansformer (SETR) [16] shows that Transformer can achieve SOTA performance in segmentation tasks as encoder. However, Transformer-based models have not attracted enough attention in medical image segmentation. TransUNet [17] utilizes CNNs to extract features and then feeds them into Transformer for long-range dependencies modeling. TransFuse [18] based on ViT tries to fuse the features extracted by Transformer and CNNs, while MedT [19] based on Axial-Attention [20] explores the feasibility of applying Transformer without large-scale datasets. The success of these models shows the great potential of Transformer in medical image segmentation, but they all only apply Transformer in encoder, which means such potential of Transformer in decoder for segmentation remains to be validated.

Moreover, multi-scale feature representations have been proved to play an important role in vision transformers. Cross-Attention Multi-Scale Vision Transformer (CrossViT) [21] proposes a novel dual-branch Transformer architecture to extract multi-scale features for image classification. Multi Vision Transformers (MViT) [22] is present for video and image recognition by connecting multi-scale feature hierarchies with transformer models. Multi-modal Multi-scale TRansformer (M2TR) [23] uses a multi-scale transformer to detect the local inconsistency at different scales. In general, multi-scale feature presentations can bring more powerful performance to vision transformers, but they are rarely used in the field of image segmentation.

To alleviate the inherent inductive biases of CNNs, this paper proposes a novel encoder-decoder Transformer based framework that mainly combines the advantages of Swin Transformer and multi-scale vision transformers to effectively optimize the structure of the standard U-shaped architecture for automatic medical image segmentation. Instead of using the traditional encoder structure, the proposed DS-TransUNet adopts dual-scale encoder subnetworks based on hierarchical Swin Transformer under the different scales of image inputs. Specifically, each medical image is first sliced into non-overlapping patches at large and small scales, respectively. By taking these two different scale patches as inputs, the proposed dual-scale encoder subnetworks can effectively extract the coarse and fine-grained feature representations of different semantic scales, respectively. To make full use of these obtained features, a robust Transformer Interactive Fusion (TIF) module is designed to aggregate the multi-scale feature representations of Swin Transformer between these two encoder subnetworks, which is the key to our DS-TransUNet method. In particular, the coarse-fine-tuning feature representations from two encoder branches will be reshaped into a token of specified size, and then fed into

the TIF module to perform an effective interaction potential with each other through the self-attention mechanism of the standard Transformer. Moreover, we also introduce the Swin Transformer block into the decoder, which helps build long-range dependencies and global context connections during up-sampling. Finally, the fused features are gradually restored to the same resolution as the input images for pixel-level predictions. Benefiting from these improvements, the proposed DS-TransUNet can effectively improve the semantic segmentation quality of medical images. We evaluate the effectiveness of DS-TransUNet across four typical tasks of medical image segmentation, covering the datasets of Polyp Segmentation, ISIC 2018, GLAnd Segmentation (GLAS), and 2018 Data Science, and the experimental results consistently demonstrate the superiorities of the proposed DS-TransUNet. The main contributions of our work are as follows:

- (1) By incorporating the advantages of hierarchical Swin Transformer into both encoder and decoder, the proposed DS-TransUNet can effectively model long-range dependencies and multi-scale context connections during the process of down-sampling and up-sampling. To the best of our knowledge, this work is might be the first attempt to combine the Swin Transformer with U-shaped architecture for automatic medical image segmentation.
- (2) We introduce dual-branch Swin Transformer to extract multi-scale feature representations in the encoder, which enables the model to effectively capture coarse-fine-tuning features of different semantic scales, improving the quality of feature learning.
- (3) The TIF module is able to establish effective global dependencies between coarse and fine-grained feature representations based on self-attention mechanism, which can guarantee the coarse-fine-tuning features of semantic consistency.
- (4) Extensive experiments across four typical tasks for medical image segmentation show that the proposed DS-TransUNet consistently outperforms previous state-of-the-art methods especially in polyp segmentation task, which demonstrates the effectiveness of our method.

II. RELATED WORK

In this section, we first summarize the most typical CNN-based methods used in medical image segmentation, then we make a overview of the recent related works about vision transformers, especially in the field of segmentation. Finally, we review the existing methods which perform multi-scale feature representations and compare these methods with our proposed method.

A. Medical image segmentation based on CNNs

Convolutional neural networks (CNNs), especially encoder-decoder based U-Net [1] and its variants have demonstrated superb performance in medical image segmentation, e.g., UNet++ [2] designs a series of nested and dense skip pathways to reduce the semantic gap, Attention U-Net [4] proposes a novel attention gate (AG) mechanism that enables the

model to focus on targets of different shapes and sizes, ResUNet [3] adds weighted attention mechanism to improve the performance of retinal vessel segmentation, DenseUNet [5] takes the advantages of dense connections and skip connection of U-Net, R2U-Net [6] combines the strengths of residual networks and U-Net to achieve better feature representation, KiU-Net [7] proposes a novel architecture utilizing both under-complete and over-complete features that makes improvement in segmenting small anatomical structures, DoubleU-Net [24] uses two U-Net in sequence and adopts Atrous Spatial Pyramid Pooling (ASPP) [25], UNet 3+ [8] leverages deep supervisions and full-scale skip connections, and feed attention network (FANet) [26] unifies the previous epoch mask with the current epoch feature map during training. Note that all these methods are still based on CNNs.

B. Vision Transformer

Inspired by the success of Transformer [12] in various NLP tasks, more and more Transformer-based methods appear in CV tasks. Among the recent vision transformers, ViT [14] is the first attempt that proves pure Transformer-based architecture can achieve SOTA performance on image recognition when pre-training on large datasets such as ImageNet-22K and JFT-300M. DeiT [27] introduces data-efficient training strategies and knowledge distillation that allow ViT to perform well on smaller ImageNet-1K dataset. Swin Transformer [15] has linear computational complexity through proposed shifted window based self-attention and achieves SOTA performance in image recognition, dense prediction tasks such as object detection and semantic segmentation. Unlike most previous Transformer-based models, Swin Transformer is a hierarchical architecture which has the flexibility to be a general-purpose backbone network. SETR [16] treats semantic segmentation as a sequence-to-sequence prediction task by using transformer as encoder. In medical image segmentation, TransUNet [17] proves that Transformer can be used as powerful encoders for medical image segmentation. TransFuse [18] is proposed to improve efficiency for global context modeling by fusing transformers and CNNs. Furthermore, to train the model effectively on medical images, MedT [19] introduces Gated Axial-Attention based on Axial-DeepLab [20]. Inspired by these approaches, we propose a UNet-like architecture which applies Swin Transformer block to both encoder and decoder. It is our belief that a unified architecture across encoder and decoder based on Transformer could provide strong performance in medical image segmentation.

C. Multi-Scale Tranformer

Multi-scale feature representations based on CNNs are a classic concept in computer vision, and have shown to benefit various CV tasks [28] [29] [30] [31]. Especially, the classic feature pyramid networks (FPN) [32] has been widely adopted in object detection and semantic segmentation. However, such benefits have not been explored much in vision transformers. The close works include: CrossViT [21] proposes a dual-branch transformer and cross-attention for image classification. M2TR [23] introduces a multi-scale transformer that

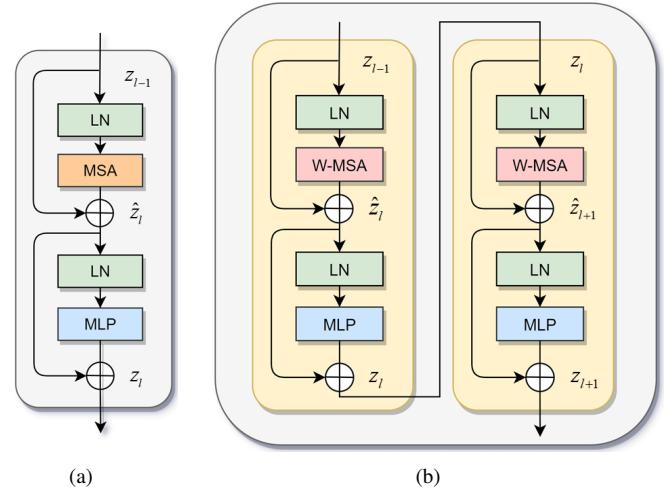


Fig. 1: (a) The architecture of a standard Transformer block (notation presented with Eq. (1)); (b) The architecture of a Swin Transformer block (notation presented with Eq. (2) and Eq. (3)).

operates on different patch sizes of feature representations. MViT [22] provides a multi-scale pyramid of features inside the transformers. Motivated by the great potential of multi-scale vision transformers, we propose a dual-branch encoder which benefits from the hierarchical architecture of Swin Transformer. Moreover, we design a efficient module called Transformer Interactive Fusion (TIF) module to fuse the multi-scale feature representations.

III. METHOD

In this section, the overall structure of proposed DS-TransUNet is introduced in detail and illustrated in Fig. 2. We first introduce the standard Transformer and Swin Transformer adopted in DS-TransUNet, then we elaborate the encoder and decoder based on Swin Transformer block since our model is a U-shaped architecture. Finally, we show that our DS-TransUNet can benefit from the dual-branch encoder design and describe how multi-scale feature representations are effectively fused by Transformer Interactive Fusion (TIF) module.

A. Swin Transformer block

The standard Transformer encoder [12] is composed of a stack of L identical blocks. As shown in 1(a), each block is consist of Multi-head Self Attention (MSA) and Multi Layer Perceptron (MLP). Besides, a LayerNorm (LN) layer is applied before each MSA module and each MLP, and a residual connection is applied after each module. Therefore the output z_l of l -layer in Transformer encoder can be expressed as:

$$\begin{aligned}\hat{z}_l &= \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \\ z_l &= \text{MLP}(\text{LN}(\hat{z}_l)) + \hat{z}_l,\end{aligned}\quad (1)$$

In the standard Transformer architecture , every token needs to be computed its relationships with all other tokens, where the computational complexity is quadratic equal to the number of tokens, making it unacceptable for many dense prediction and high-resolution image tasks. For efficient modeling, Swin

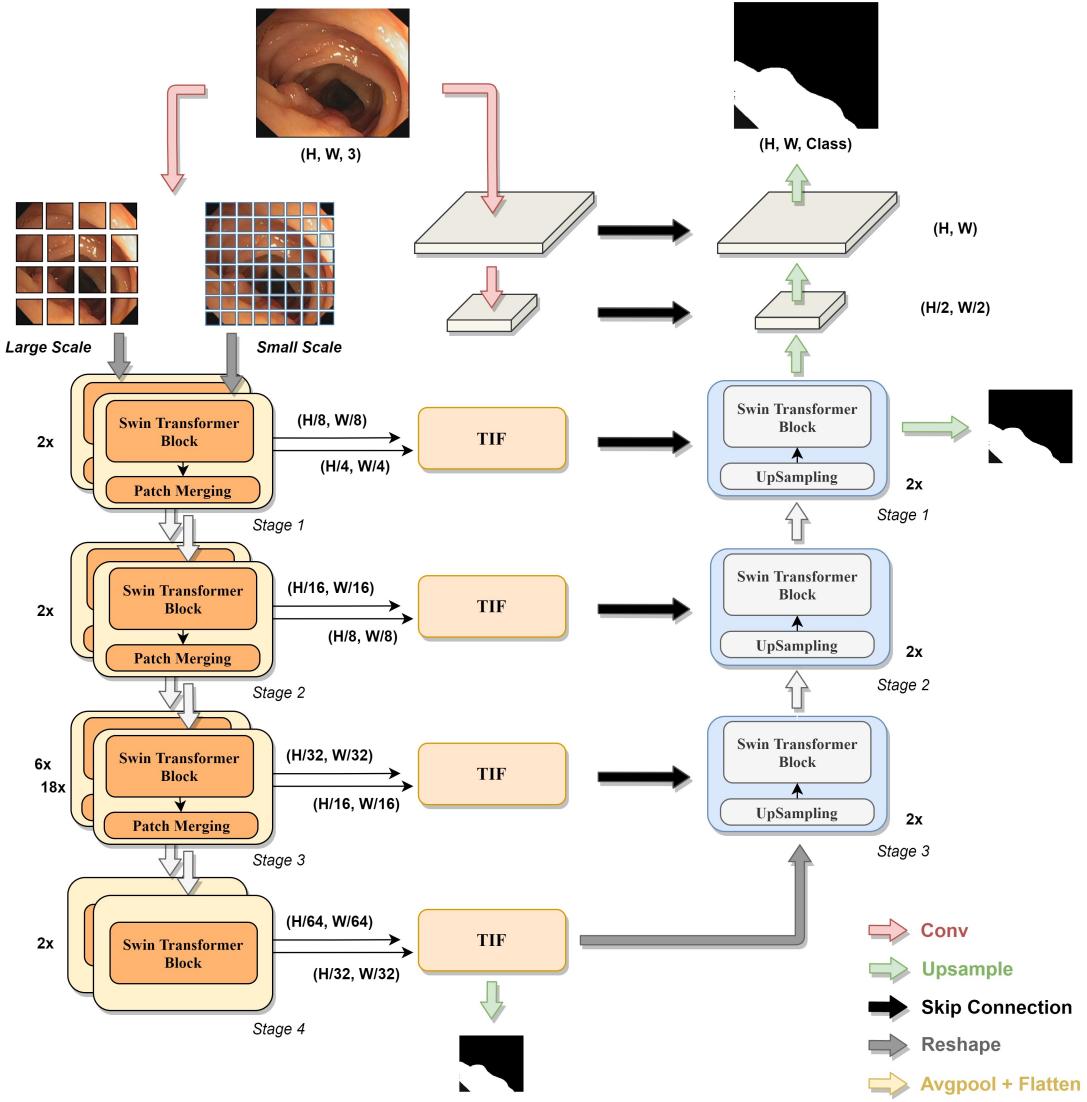


Fig. 2: Illustration of the proposed Dual Swin Transformer U-Net (DS-TransUNet). Given an input medical image, we first split it into non-overlapping patches at two scales and feed them into the two branches of encoder separately, then the output feature representations of different scales will be fused by Transformer Inter-active Fusion (TIF) module. Finally, the fused features are restored to the same resolution as input image after the up-sampling process based on Swin Transformer block, hence obtaining the final mask predictions.

Trasnformer [15] proposes Window based MSA (W-MSA) and Shifted Window based MSA (SW-MSA).

In W-MSA, the input feature will be divided into non-overlapping windows, and each window contains $M \times M$ patches (set to 7 by default). W-MSA will only conduct self-attention within local windows. As shown in Fig. 1(b), \hat{z}_l and z_l represent the outputs of W-MSA and MLP in l^{th} layer , which are computed as:

$$\begin{aligned} \hat{z}_l &= \text{W-MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \\ z_l &= \text{MLP}(\text{LN}(\hat{z}_l)) + \hat{z}_l, \end{aligned} \quad (2)$$

The problem of W-MSA is the lack of effective information interaction between windows, in order to introduce cross-window interaction without additional computation, there exists a SW-MSA followed by the W-MSA.

The window configuration of SW-MSA is different from the previous W-MSA layer where it proposes an efficient batch

processing method by cyclic-shifting to the upper-left. After this shift, a batch window may be consisted of multiple non-adjacent sub-windows in the feature map and keep the equal number of batch windows as regular partitioning at the same time. While conducting self-attention within local windows in both W-MSA and SW-MSA, the relative position bias is included in computing similarity.

With such shifted window partitioning mechanism, the outputs of SW-MSA and MLP module can be written as:

$$\begin{aligned} \hat{z}_{l+1} &= \text{SW-MSA}(\text{LN}(z_l)) + z_l, \\ z_{l+1} &= \text{MLP}(\text{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1}, \end{aligned} \quad (3)$$

B. Encoder

In the overall structure of our model, we refer to [1] using the U-shaped architecture. For encoder, the Swin Transformer [15] is used for feature extraction. As shown in Fig. 2, the

input medical image will first be sliced into $\frac{H}{s} \times \frac{H}{s}$ non-overlapping patches, where s is the patch size. Each patch is treated as a “token” and will be projected to dimension C by linear embedding layer. Since the patches are obtained by convolution operation, no additional position information is needed here. These patch tokens are formally fed into Swin Transformer, which contains four stages, and each stage holds a certain number of Swin Transformer blocks that include window multi-head self attention (W-MSA) and shifted window multi-head self attention (SW-MSA). To produce a hierarchical representation, the number of tokens will be reduced as the network gets deeper; in the first three stages, input features will go through patch merging layer to reduce the feature resolution and increase dimension after Swin Transformer blocks’ transformation. Specifically, the patch merging layer concatenates features of each group of 2×2 neighboring patches, and then applies a linear layer on the channel-dimensional concatenated features. This will reduce the number of tokens by $2 \times 2 = 4$, $2 \times$ downsampling of resolution and increase the output dimension by 2. So the output resolutions of four stages are $\frac{H}{s} \times \frac{H}{s}$, $\frac{H}{2s} \times \frac{H}{2s}$, $\frac{H}{4s} \times \frac{H}{4s}$ and $\frac{H}{8s} \times \frac{H}{8s}$; and the dimensions are C , $2C$, $4C$ and $8C$ respectively.

C. Decoder

As shown in Fig. 2, the decoder mainly consists of three stages. Unlike the previous U-Net [1] and its variants, each stage of our model includes not only up-sampling (Nearest Upsampling) and skip connection, but also Swin Transformer block. Specifically, the output of stage 4 in encoder is used as initial input of decoder. In each stage of decoder, the input features are up-sampled by 2, and then concatenated with the appropriate skip connection feature maps from encoder in the same stage. After that, the output is fed into Swin Transformer block. We choose this design since 1) it allows us to make full use of the features from encoder and up-sampling 2) it can build long-range dependencies and global context interaction in decoder to achieve better decoding performance. The impact of introducing Swin Transformer block in decoder will be discussed in section V-B.

After the three stages above, we can get the output with resolution of $\frac{H}{4} \times \frac{H}{4}$. Using a $4 \times$ upsampling operator directly will lost a lot of shallow features, so we down-sampling the input image by cascading two blocks to get the low level feature with resolution of $H \times W$ and $\frac{H}{2} \times \frac{H}{2}$, where each block consists a 3×3 convolutional layer, a group normalization layer and a ReLU layer successively. All these output features will be used to get the final mask predictions through skip connection.

D. Multi-Scale Feature Representations

Although self-attention can effectively build long-range dependencies between patches, patch division ignores the pixel-level intrinsic structure features inside each patch, which will lead to the lose of shallow features such as edges and lines information. Moreover, ViT [14] can obtain better performance with fine-grained patch size. Taking these into account, and in

order to improve the segmentation performance and enhance the robustness of our model, we employ multi-scale Swin Transformer for feature extraction.

Patches of different scales can complement each other in feature extraction; the large scale can better capture coarse-grained feature, while small patch can better obtain the fine-grained feature. Although the convolutional layer can introduce location information between patches implicitly, the information is lost at pixel level within each patch. In [21], dual-branch Transformer can alleviate the above problems to a certain extent, and achieve better performance than ViT in image recognition. Motivated by this, we propose multi-scale Swin Transformer in encoder. More specifically, we use two independent branch with patch size of $s = 4$ (primary) and $s = 8$ (complementary) for feature extraction at different spatial levels. As result, the output with resolutions of $\frac{H}{4} \times \frac{H}{4}$, $\frac{H}{8} \times \frac{H}{8}$, $\frac{H}{16} \times \frac{H}{16}$ and $\frac{H}{32} \times \frac{H}{32}$ can be obtained from small-scale branch, while output resolutions of large-scale are $\frac{H}{8} \times \frac{H}{8}$, $\frac{H}{16} \times \frac{H}{16}$, $\frac{H}{32} \times \frac{H}{32}$ and $\frac{H}{64} \times \frac{H}{64}$.

E. Transformer Interactive Fusion Module (TIF)

After obtaining the output features from dual-branch encoder, the remaining problem is how to fuse them since effective feature fusion is the core of multi-scale feature representations learning. A direct approach is to simply concatenate the multi-scale features and then perform convolution operation. However, such straightforward approach fails to capture the long-range dependencies and global context connection between features at different scales. Therefore, we propose a novel Transformer Interactive Fusion (TIF) module, which utilizes the MSA mechanism to enable efficient and effective interaction between multi-scale features. In particular, we select the standard Transformer block [12] instead of Swin Transformer block in TIF, mainly because the latter essentially operates on rectangle-based feature map, while in multi-scale features fusion module, we need to generate a token at specified size based on feature map of one branch, and then compute self-attention together with the token sequence reshaped by another branch. Moreover, we only need to perform monolayer self-attention operation twice at each stage, which means the computational complexity is acceptable.

As shown in Fig. 3, the proposed TIF can integrate features from two branches of different scales. In the following, we choose the small scale branch for specific analysis, and the same procedure is also applicable to large scale branch.

To be specific, for outputs of two branches from the same stage i ($i = 1, 2, 3, 4$) denoted as $F^i = [f_1^i, f_2^i, \dots, f_{h \times w}^i] \in \mathbb{R}^{C \times (h \times w)}$ (primary branch) and $G^i = [g_1^i, g_2^i, \dots, g_{\frac{h}{2} \times \frac{w}{2}}^i] \in \mathbb{R}^{c \times (\frac{h}{2} \times \frac{w}{2})}$ (complementary branch), respectively. Then we obtain the transformation output of G^i by:

$$\hat{g}^i = \text{Flatten}(\text{Avgpool}(G^i)), \quad (4)$$

where $\hat{g}^i \in \mathbb{R}^{C \times 1}$, Avgpool is a 1 dimension average pooling layer, followed by flatten operation. The token \hat{g}^i represents the global abstract information of G^i to interact with F^i at pixel level. Meanwhile, F^i is concatenated with \hat{g}^i into a

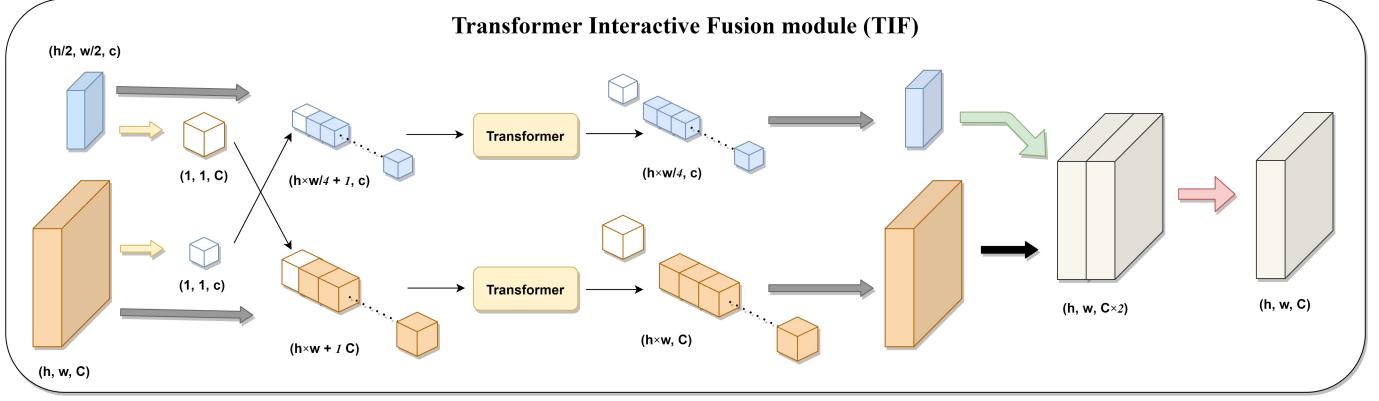


Fig. 3: Illustration of Transformer Interactive Fusion module (TIF), which serves as the core component of DS-TransUNet in the multi-scale features fusion process.

sequence of $1 + h \times w$ tokens, which is fed into Transformer layer for computing global self-attention:

$$\begin{aligned} \hat{F}^i &= \text{Transformer}([\hat{g}^i, f_1^i, f_2^i, \dots, f_{h \times w}^i]), \\ &= [\hat{f}_0^i, \hat{f}_1^i, \dots, \hat{f}_{h \times w}^i] \in \mathbb{R}^{C \times (1+h \times w)} \\ F_{out}^i &= [\hat{f}_1^i, \hat{f}_2^i, \dots, \hat{f}_{h \times w}^i] \in \mathbb{R}^{C \times (h \times w)} \end{aligned} \quad (5)$$

where *Transformer* plays the same role as Eq. 1 and F_{out}^i as the final output of small scale branch in TIF. This approach introduces connections between each token in $F^i = [f_1^i, f_2^i, \dots, f_{h \times w}^i] \in \mathbb{R}^{C \times (h \times w)}$ and the whole G^i , so that fine-grained feature can also obtain coarse-grained information from the large scale branch. Therefore, the TIF module can bring effective feature fusion of multi-scale branch which helps achieve better segmentation performance. The impact of TIF compared to ordinary multi-scale features fusion based on CNN will be discussed in section V-B.

IV. EXPERIMENTS

To evaluate the learning and generalization ability of our Dual Swin Transformer U-Net (DS-TransUNet), we conduct experiments on four common medical image segmentation tasks with several publicly available datasets, and compare them with other SOTA methods. In this section we present the basic information about all the datasets briefly. Besides, we also describe the evaluation metrics and implementation details.

A. Datasets

Polyp Segmentation: For polyp segmentation task, we select five public polyp datasets including Kvasir [33], CVC-ColonDB (ColonDB) [34], EndoScene [35], ETIS [36], and CVC-ClinicDB (ClinicDB) [37]. The split and training settings of these datasets are different in [38], [24] and [39], so we conduct experiments according to these three articles respectively.

- In [39], only Kvasir is used, which contains 880 images for training and 220 images for testing. For this split, we resize each image to a resolution of 512×512 .

- According to [24], we only use ClinicDB during experiment, of which 550 images are used for training while 62 for testing. Besides, all the images used are resized to 384×384 .
- As for [38], the training sets consist 900 images in Kvasir and 550 images in ClinicDB, while the testing sets contain five datasets, which are Kvasir with 100 images, ClinicDB with 62 images, ColonDB with 380 images, EndoScene with 60 images and ETIS with 196 images. Since the resolutions of images in datasets are not uniform, we resize them to 384×384 .

ISIC 2018 Dataset: The dataset comes from ISIC-2018 challenge [40] [41] and is useful for skin lesion analysis. It includes 2596 images and their corresponding annotations, which are resized to 256×256 . The images are randomly split into 2076 images for training and 520 images for testing. This process is repeated five times and the average is taken as result.

GLAS Dataset: GLAnd Segmentation (GLAS) dataset is from 2015 challenge on gland segmentation in histology images, which provides images of Haematoxylin and Eosin (H&E) stained slides. It contains 165 images which are split into 85 images for training and 80 for testing according to [7]. Besides, images are resized into 128×128 .

2018 Data Science Bowl: The dataset is from 2018 Data Science Bowl challenge [42] and used to find the nuclei in divergent images, including 670 images in total. We use the same settings as [38], 80% of dataset for training, 10% for validation, and 10% for testing. Moreover, all the images are resized into 256×256 .

B. Evaluation Metrics

To compare SOTA methods with our proposed DS-TransUNet, the standard evaluation metrics that we use include mean Dice Coefficient (mDice) (a.k.a. F1), mean Intersection over Union (mIoU), precision and recall, which are associated with four values i.e., true-positive (TP), true-negative (TN),

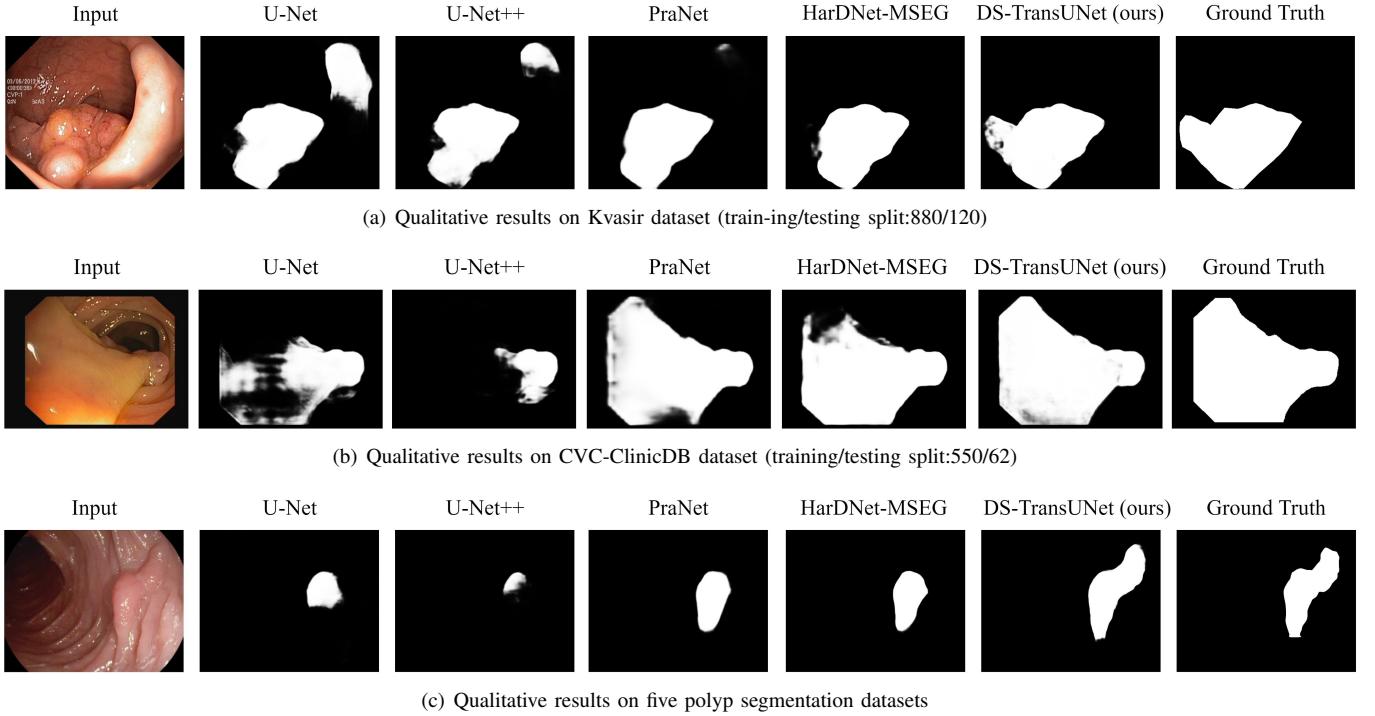


Fig. 4: Qualitative results on polyp segmentation task of DS-TransUNet compared to other models. Our model shows better learning and generalization ability, which leads to higher-quality segmentation performance.

false-positive (FP), and false-negative (FN).

$$\begin{aligned}
 mDice &= \frac{2 \times TP}{2 \times TP + FP + FN}, \\
 mIoU &= \frac{TP}{TP + FP + FN}, \\
 Precision &= \frac{TP}{TP + FP}, \\
 Recall &= \frac{TP}{TP + FN},
 \end{aligned} \tag{6}$$

C. Implementation Details

Multi-scale training strategy is used in all experiments instead of data augmentation. The loss functions used are weighted IoU loss \mathcal{L}_{IoU}^W and binary cross-entropy loss \mathcal{L}_{BCE}^W . Inspired by [38], we find deep supervision helps the model training by additionally supervising the output S_2 of stage 4 in encoder and S_3 of stage 1 in decoder, which means the final loss function \mathcal{L}_{total} can be written as:

$$\begin{aligned}
 \mathcal{L}_{total} &= \alpha \mathcal{L}(G, S_1) + \beta \mathcal{L}(G, S_2) + \gamma \mathcal{L}(G, S_3), \\
 \mathcal{L} &= \mathcal{L}_{IoU}^W + \mathcal{L}_{BCE}^W,
 \end{aligned} \tag{7}$$

where G is the groundtruth in training sample and α, β, γ are hyperparameters which are set to 0.6, 0.2, 0.2 empirically. We train our model with SGD optimizer with momentum 0.9, weight decay 1e-4 and learning rate equals to 0.01.

All models are trained for 100 epochs. Moreover, early stopping and Cosine Annealing schedule are also used. All models are built using PyTorch framework and trained on a NVIDIA RTX 3090 GPU. Our model is provided in two variants: the base version (DS-TransUNet-B) uses Swin-Base

[15] as primary scale branch (small scale branch) for encoder, while the large version (DS-TransUNet-L) uses Swin-Large [15]. Both the two version use Swin-Tiny [15] as complementary scale branch (large scale branch) for encoder. All these sizes of Swin Transformer use pretrained weights released from [15]. The detailed parameters of model are summarized in Table I, where Layer Number and Head Number mean the number of Swin Transformer block and head self-attention in each stage respectively. Moreover, Window Size represents the size of non-overlapping windows divided in W-MSA, and Swin-Decoder refers to the Swin Transformer block used in decoder.

V. RESULT

In this section, we conduct experiments to compare our proposed model with SOTA methods on four segmentations tasks. Besides, we also present the experimental results and visualize some qualitative results to evaluate the learning and generalization ability of our DS-TransUNet. Finally, we also perform ablation study on polyp segmentation task to analyze the effect of each proposed technique used in DS-TransUNet.

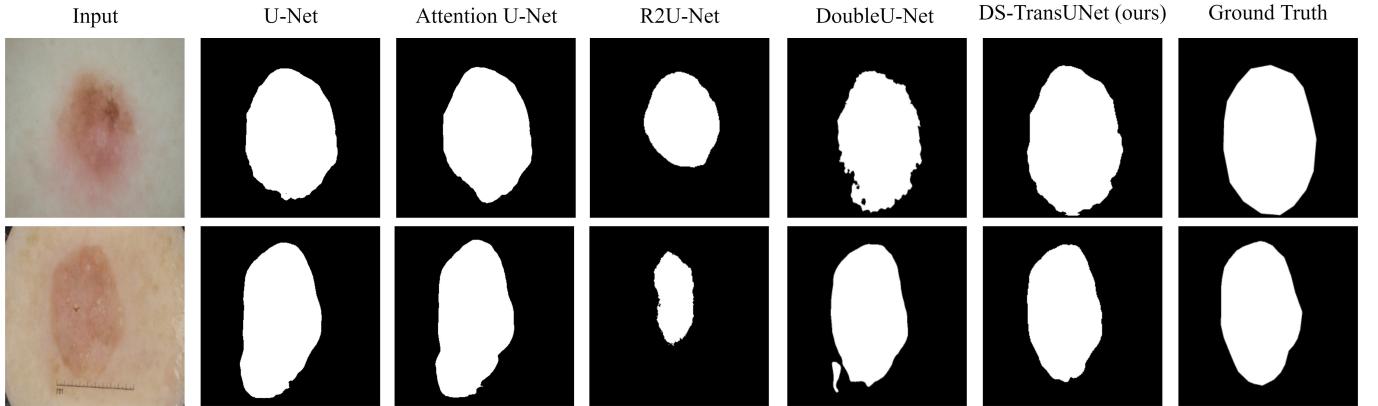
A. Comparison with State-of-the-art Methods

Results on Polyp Segmentation: Our quantitative results on polyp segmentation task achieve SOTA performance compared to other models, which is present in Table II, III and V. Next, we analyze the quantitative results on the three kinds of data splits.

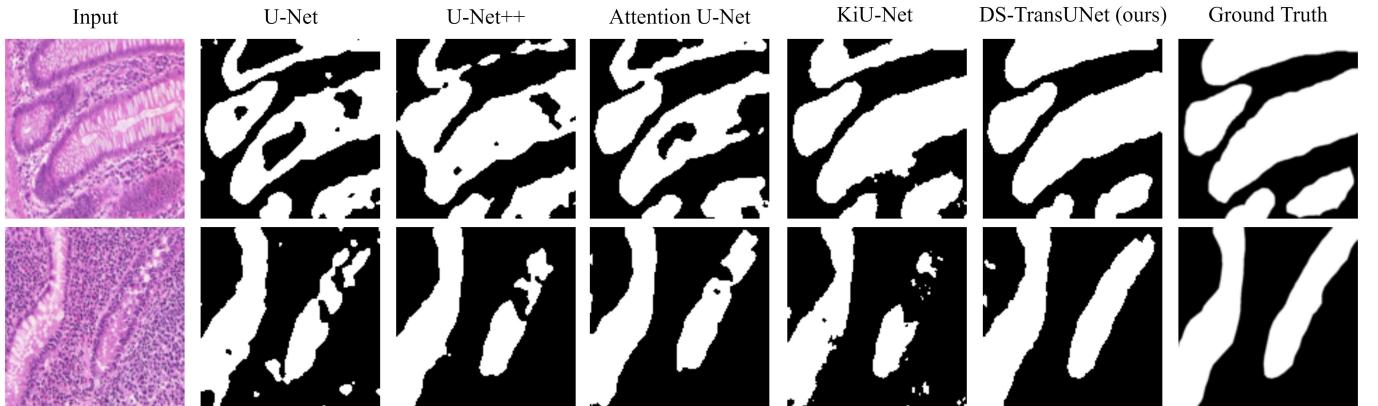
In [39], only Kvasi dataset [33] is used and the evaluation metrics used include mDice, mIoU, recall and precision. From

TABLE I: DETAILS OF SWIN TRANSFORMER MODEL VARIANTS.

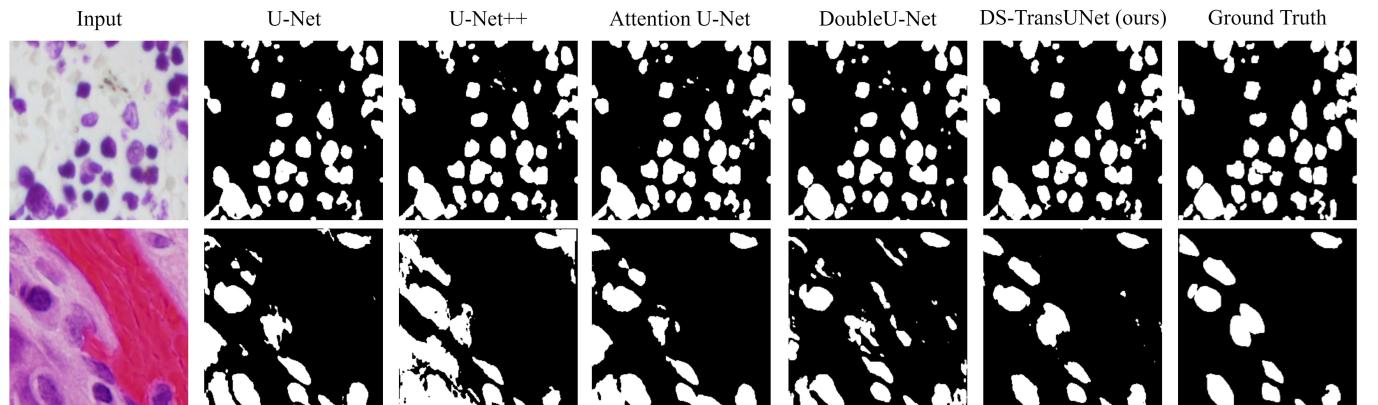
Methods	Hidden Size C	MLP Size D	Layer Number	Head Number	Window Size
Swin-Tiny	96	384	[2, 2, 6, 2]	[3, 6, 12, 24]	7
Swin-Base	128	512	[2, 2, 18, 2]	[4, 8, 16, 32]	7
Swin-Large	192	768	[2, 2, 18, 2]	[6, 12, 24, 48]	7
Swin-Decoder	128	512	[2, 2, 2]	[8, 4, 2]	7



(a) Quantitative results on ISIC 2018 dataset



(b) Quantitative results on GLAS dataset



(c) Quantitative results on 2018 Data Science Bowl dataset

Fig. 5: Qualitative results of DS-TransUNet on three medical image segmentation tasks compared to other models.

TABLE II: QUANTITATIVE RESULTS ON KVASIR DATASET (TRAINING/TESTING SPLIT:880/120). FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	mDice	mIoU	Recall	Precision
U-Net [1]	0.597	0.471	0.617	0.672
Res-UNet [3]	0.690	0.572	0.725	0.745
ResUNet++ [43]	0.714	0.613	0.742	0.784
DoubleU-Net [24]	0.813	0.733	0.840	0.861
FCN8 [44]	0.831	0.737	0.835	0.882
PSPNet [45]	0.841	0.744	0.836	0.890
HRNet [46]	0.845	0.759	0.859	0.878
DeepLabv3+ [47]	0.864	0.786	0.859	0.906
FANet [26]	0.880	0.810	0.906	0.901
HarDNet-MSEG [48]	0.904	0.848	0.923	0.907
DS-TransUNet-B (ours)	0.911	0.856	0.935	0.914
DS-TransUNet-L (ours)	0.913	0.859	0.936	0.916

TABLE III: QUANTITATIVE RESULTS ON CVC-CLINICDB DATASET (TRAINING/TESTING SPLIT:550/62). FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	F1	mIoU	Recall	Precision
SFA [49]	0.7000	0.6070	-	-
ResUNet-mod [50]	0.7788	0.4545	0.6683	0.8877
UNet++ [2]	0.7940	0.7290	-	-
ResUNet++ [24]	0.7955	0.7962	0.7022	0.8785
U-Net [1]	0.8230	0.7550	-	-
PraNet [38]	0.8990	0.8490	-	-
DoubleU-Net [24]	0.9239	0.8611	0.8457	0.9592
FANet [26]	0.9355	0.8937	0.9339	0.9401
DS-TransUNet-B (ours)	0.9350	0.8845	0.9464	0.9306
DS-TransUNet-L (ours)	0.9422	0.8939	0.9500	0.9369

Table II, we can see that not only DS-TransUNet-L, but also DS-TransUNet-B outperforms the previous SOTA HarDNet-MSE [48] on all metrics. Specifically, DS-TransUNet-L achieves a mDice of 0.913, mIoU of 0.859, recall of 0.936 and precision of 0.916 with an improvement of 0.9%, 1.1%, 1.3% and 0.9%. As visualized in Fig. 4(a), our DS-TransUNet achieve the best segmentation performance among all models, especially for fuzzy polyps at the edge of image, which are often missed-out in colonoscopy because their color and structure are similar to the surrounding intestinal tissue.

According to [24], only CVC-ClinicDB dataset is used during experiment. Table III shows that our proposed DS-TransUNet-L achieves SOTA results on almost all metrics (F1, mIoU, and recall). Specifically, DS-TransUNet-L outperforms the previous SOTA FANet [26] by an improvement of 0.67%, 0.02% and 1.6% in terms of F1, mIoU and recall, while produces a comparable precision score compared to DoubleU-Net [24]. From Fig. 4(b), we can see that polyps with large area can also be accurately segmented. Moreover, the higher recall score also shows that our DS-TransUNet is more clinically useful.

TABLE IV: QUANTITATIVE RESULTS ON ISIC 2018 DATASET. FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	F1	mIoU	Recall	Precision
U-Net [1]	0.6740	0.5490	0.7080	-
Attention U-Net [4]	0.6650	0.5660	0.7170	-
R2U-Net [6]	0.6790	0.5810	0.7920	-
Attention R2U-Net [6]	0.6910	0.5920	0.7260	-
BCDU-Net (d=3) [51]	0.8510	-	0.7850	-
FANet [26]	0.8731	0.8023	0.8650	0.9235
DoubleU-Net [24]	0.8962	0.8212	0.8780	0.9459
DS-TransUNet-B (ours)	0.9101	0.8481	0.9108	0.9337
DS-TransUNet-L (ours)	0.9132	0.8523	0.9217	0.9271

Referring to [38], where the training sets are consist of Kvasir and CVC-ClinicDB, and testing sets additionally include three unseen datasets. As for quantitative evaluation, we use mDice and mIoU following [38]. Our proposed model achieves SOTA performance on all five challenging dataset, which is the only model that produces over 0.92 and 0.93 mDice on Kvasir. In particular, DS-TransUNet-B and DS-TransUNet-L both outperform the latest TransFuse [18] on the two in-domain datasets. As for unseen datasets (ColonDB, EndoSene and ETIS), our DS-TransUNet also greatly exceeds all SOTA methods by a large margin. Specifically, DS-TransUNet-B achieves better performance on all datasets except EndoScene, while DS-TransUNet-L even yields the top performance on all five datasets. In general, our proposed method outperforms SOTA methods with mDice improvement of 1.7%, 0.4%, 2.5%, 0.7% and 3.5%, and achieves about 1.9% improvement in terms of the average mDice score compared with TransFuse, which shows the advantages and strong learning ability of our proposed model. The qualitative segmentation performance in Fig. 4(c) also shows the great generalization ability of DS-TransUNet.

Results on ISIC 2018 Dataset: For 2018 ISIC dataset, the metrics used are F1 Score, mIoU, recall and precision. Table IV presents the specific results, where our proposed model achieves better segmentation performances than SOTA DoubleU-Net [24] and the latest FANet [26]. DS-TransUNet-L produces 0.9132 on F1, 0.8523 on mIoU and recall of 0.9217 with an improvement of 1.70%, 3.11% and 4.37% compared with SOTA method, respectively. Although DoubleU-Net outperforms in terms of precision, our proposed model produces the best overall performances on four metrics. As shown in Fig. 5(a), the qualitative results qualitatively manifest that our proposed method can not only accurately predict the location and boundary of skin lesion, but also better distinguish it from normal skin.

Results on GLAS Dataset: Our quantitative results on the GLAS dataset are shown in Table VI. Comparing with the leading SOTA method KiU-Net [7], our proposed methods DS-TransUNet-B and DS-TransUNet-L both outperform KiU-Net on both mDice and mIoU. Especially DS-TransUNet-L achieves a 3.94% improvement in terms of mDice and 5.67%

TABLE V: QUANTITATIVE RESULTS ON FIVE POLYP SEGMENTATION DATASETS COMPARED TO PREVIOUS SOTA METHODS. FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

	Kvasir		ClinicDB		ColonDB		EndoScene		ETIS		Average	
	mDice	mIoU										
U-Net [1]	0.818	0.746	0.823	0.755	0.512	0.444	0.398	0.335	0.710	0.626	0.652	0.581
U-Net++ [2]	0.821	0.743	0.794	0.729	0.483	0.410	0.401	0.344	0.707	0.624	0.641	0.570
PraNet [38]	0.898	0.840	0.899	0.849	0.709	0.640	0.871	0.797	0.628	0.567	0.800	0.739
HarDNet-MSEG [48]	0.912	0.857	0.932	0.882	0.731	0.660	0.887	0.821	0.677	0.613	0.828	0.767
TransFuse-S [18]	0.918	0.868	0.918	0.868	0.773	0.696	0.902	0.833	0.733	0.659	0.849	0.785
TransFuse-L [18]	0.918	0.868	0.934	0.886	0.744	0.676	0.904	0.838	0.737	0.661	0.847	0.786
DS-TransUNet-B (ours)	0.934	0.888	0.938	0.891	0.798	0.717	0.882	0.810	0.772	0.698	0.865	0.801
DS-TransUNet-L (ours)	0.935	0.889	0.936	0.887	0.798	0.722	0.911	0.846	0.761	0.687	0.868	0.806

TABLE VI: QUANTITATIVE RESULTS ON THE GLAS DATASET. FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	mDice	mIoU
Seg-Net [52]	78.61	65.96
U-Net [1]	79.76	67.63
MedT [19]	81.02	69.61
UNet++ [2]	81.13	69.61
Attention UNet [4]	81.59	70.06
KiU-Net [7]	83.25	72.78
DS-TransUNet-B (ours)	86.54	77.36
DS-TransUNet-L (ours)	87.19	78.45

TABLE VII: QUANTITATIVE RESULTS ON THE 2018 DATA SCIENCE BOWL. FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	F1	mIoU	Recall	Precision
U-Net [1]	0.7573	0.9103	-	-
UNet++ [2]	0.8974	0.9255	-	-
Attention UNet [4]	0.9083	0.9103	-	0.9161
DoubleU-Net [24]	0.9133	0.8407	0.6407	0.9496
FANet [26]	0.9176	0.8569	0.9222	0.9194
DS-TransUNet-B (ours)	0.9200	0.8589	0.9427	0.9054
DS-TransUNet-L (ours)	0.9219	0.8612	0.9378	0.9124

of mIoU over SOTA method. GLAS is a dataset with only 85 training samples, and we can achieve SOTA performance only by using multi-scale training, which effectively proves that our proposed method can produce high-quality segmentation performance even on a small-scale datasets. Besides, we also present the visualization of generated mask images in Fig. 5(b), which demonstrates that our model is able to better distinguish the gland itself from the surrounding tissue, and bring excellent gland segmentation performance.

Results on 2018 Data Science Bowl: For 2018 data science bowl challenge, we compare our result with the SOTA models. Table VII shows that DS-TransUNet-L achieves a F1 of 0.9219, mIoU of 0.8612 and recall of 0.9378, which are

0.43%, 0.43% and 1.56% higher than the best performing FANet [26]. Besides, DS-TransUNet-B can yield the highest recall score of 0.9427. In general, although UNet++ and DoubleU-Net still keep the SOTA performance in mIoU (0.9255) and precision (0.9496) respectively, our proposed model achieves the best balance among the four metrics compared to the other SOTA methods. From the qualitative results in Fig. 5(c), we can observe that our DS-TransUNet can better capture the presence of cell nuclei and bring better segmentation prediction.

B. Ablation study

In order to evaluate the ability of Swin Transformer in medical image segmentation and the influence of various factors on our proposed model, we further conduct ablation studies on four variants of our DS-TransUNet. The datasets we select are based on polyp segmentation task, which can verify the learning and generalization ability of models.

- **Base model**, which directly processes the final output of Swin Transformer by a progressive upsampling strategy. Specifically, the output in stage 4 of Swin Transformer is up-sampled by cascading three blocks, where each block consists convolution layers and $2 \times$ upsampling operations. After that, we obtain the output with resolution of $\frac{H}{4} \times \frac{W}{4}$, and then perform the same processing as described in III-C to make the final pixel-level predictions.
- **Swin U-Net**, which is based on U-shape architecture. It uses Swin Transformer as encoder, while keeps the same structure in decoder as [1], which only utilizes convolution layer, $2 \times$ upsampling and skip connection.
- **Swin Decoder**, which is based on Swin U-Net, further adding Swin Transformer block for long-range dependencies modeling after each up-sampling process. The specific Swin Transformer block parameters used in decoder are shown in Table I.
- **Multi-Scale SD**, whose full name is Multi-Scale Swin Decoder, leverages dual-branch Swin Transformer for feature extraction in encoder based on Swin Encoder. Compared with DS-TransUNet, it utilizes convolution layer for multi-scale feature representations fusion instead of TIF.

TABLE VIII: BLATION STUDY ON POLYP SEGMENTATION TASK. FOR EACH COLUMN, THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

	Kvasir		ClinicDB		ColonDB		EndoScene		ETIS		Average	
	mDice	mIoU										
TransFuse-S	0.918	0.868	0.918	0.868	0.773	0.696	0.902	0.833	0.733	0.659	0.849	0.785
TransFuse-L	0.918	0.868	0.934	0.886	0.744	0.676	0.904	0.838	0.737	0.661	0.847	0.786
Base model (Base)	0.919	0.863	0.915	0.861	0.747	0.657	0.878	0.804	0.722	0.635	0.836	0.764
Base model (Large)	0.922	0.867	0.914	0.861	0.786	0.698	0.884	0.812	0.735	0.652	0.848	0.778
Swin U-Net (Base)	0.920	0.868	0.914	0.862	0.758	0.674	0.884	0.811	0.711	0.636	0.837	0.770
Swin U-Net (large)	0.926	0.876	0.923	0.875	0.791	0.709	0.889	0.816	0.734	0.650	0.853	0.785
Swin Decoder (Base)	0.927	0.877	0.936	0.889	0.785	0.697	0.886	0.813	0.741	0.666	0.855	0.788
Swin Decoder (Large)	0.929	0.879	0.929	0.880	0.798	0.717	0.904	0.836	0.759	0.677	0.864	0.798
Multi-Scale SD (Base)	0.931	0.882	0.927	0.878	0.784	0.704	0.864	0.789	0.716	0.632	0.844	0.777
Multi-Scale SD (Large)	0.927	0.876	0.928	0.877	0.786	0.707	0.862	0.785	0.737	0.655	0.848	0.780
DS-TransUNet-B	0.934	0.888	0.938	0.891	0.798	0.717	0.882	0.810	0.772	0.698	0.865	0.801
DS-TransUNet-L	0.935	0.889	0.936	0.887	0.798	0.722	0.911	0.846	0.761	0.687	0.868	0.806

Table **VIII** presents the experimental results of four variants of DS-TransUNet on polyp segmentation task, in terms of both mean Dice and mean IoU. Moreover, we select the latest TransFuse [18] as baseline.

Effect of Swin Transformer: Swin Transformer block is the core component of our proposed method, which computes representation with W-MSA and SW-MSA, and surpasses the previous SOTA methods in multiple CV tasks. To explore the feature extraction ability of Swin Transformer in medical image segmentation task, we compare Base model with the previous SOTA TransFuse. In Table **VIII** we can see that Swin Transformer achieves satisfied segmentation performance as encoder. Although it is not as good as TransFuse in overall performance, it still produces close and comparable results. Especially Base model (Large), shows an improvement of 0.4% and 1.3% in Kvasir and ColonDB in terms of mDice respectively compared to the best results of TransFuse.

Effect of Swin Transformer block in decoder: In order to explore the influence of Swin Transformer in decoder, we conduct the experiments of two specially designed models based on Swin Transformer as encoder: Swin U-Net and Swin Decoder. The specific results shown in Table **VIII** indicate that the U-shaped encoder-decoder based architecture can effectively improve the segmentation performance. Especially Swin U-Net (Large), has achieved 0.4% improvement in terms of the average mean Dice score compared to TransFuse.

By simply adding Swin Transformer block after each up-sampling in Swin U-Net, Swin Decoder can easily build long-range dependencies and global context connection in decoder. As shown in Table **VIII**, we can see that Swin Decoder already achieves better performance than the latest TransFuse on all five challenging datasets with an improvement of 1.5% in terms of the average mDice score, which means that Swin Decoder has better learning and generalization ability than previous SOTA methods. Specially, the best results of Swin Decoder outperform TransFuse with mDice improvement of 1.1%, 0.2%, 2.1% and 2.2% in all dataset except EndoScene.

Therefore, the decoder design based on Swin Transformer block can effectively improve the segmentation performance.

Effect of multi-scale feature representations and TIF: Multi-Scale SD adds another Swin Transformer branch in encoder, and simply fuses the multi-scale features through convolution operation. Such a straightforward approach does not bring performance improvements. The experimental results shown in Table **VIII** indicates that despite the additional encoder branch is added which brings more granular information, it fails to achieve better performance compared to Swin Encoder with single branch. This is mainly because common convolution layer can not effectively fuse multi-scale feature, but makes the model more difficult to converge

By adding TIF module to Multi-scale SD, we can get the final proposed DS-TransUNet, which yields the best performance among all variants. To evaluate the effectiveness of the proposed TIF module, we compare DS-TransUNet with Swin Decoder in Table **VIII**. It can be observed that the best results of DS-TransUNet achieve mDice improvements of 0.6%, 0.2%, 0.7%, 1.3% on Kvasir, ClinicDB, EndoScene and ETIS compared to Swin Decoder with single branch, while show a 0.5% improvement on ColonDB in terms of mIoU. In general, TIF allows more efficient interaction between features of different scales, which brings more effective feature representations fusion of multi-scale branches and helps achieve better segmentation performance.

VI. CONCLUSION

In this work we present the Dual Swin Transformer U-Net (DS-TransUNet), a U-shaped encoder-decoder based framework for medical image segmentation. Our DS-TransUNet is based on the hierarchical Swin Transformer. Not only the encoder, we also innovatively add Swin Transformer block in decoder. Moreover, we introduce dual-branch Swin Transformer in encoder to extract multi-scale feature representations. We further propose a novel Transformer Interactive Fusion (TIF) module to build long-range dependencies between features of

different scales through self-attention mechanism, thus effectively fusing the multi-scale features from encoder. Extensive experiments on four medical image segmentation tasks show that our DS-TransUNet significantly outperforms other state-of-the-art methods especially in polyp segmentation task. In the future, our work will focus on designing more lightweight Transformer-based models and better learning the pixel-level intrinsic structural features generated by the patch division in vision transformers.

REFERENCES

- [1] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [2] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3–11.
- [3] 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)*. IEEE, 2018, pp. 327–331.
- [4] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [5] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, “H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes,” *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.
- [6] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, “Recurrent residual convolutional neural network based on u-net (r2unet) for medical image segmentation,” *arXiv preprint arXiv:1802.06955*, 2018.
- [7] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, “Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 363–373.
- [8] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, “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)*. IEEE, 2020, pp. 1055–1059.
- [9] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
- [10] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “Ccnet: Criss-cross attention for semantic segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612.
- [11] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” *arXiv preprint arXiv:2103.02907*, 2021.
- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.
- [13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.
- [14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” *arXiv preprint arXiv:2103.14030*, 2021.
- [16] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” *arXiv preprint arXiv:2012.15840*, 2020.
- [17] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.
- [18] Y. Zhang, H. Liu, and Q. Hu, “Transfuse: Fusing transformers and cnns for medical image segmentation,” *arXiv preprint arXiv:2102.08005*, 2021.
- [19] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, “Medical transformer: Gated axial-attention for medical image segmentation,” *arXiv preprint arXiv:2102.10662*, 2021.
- [20] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, “Axial-deeplab: Stand-alone axial-attention for panoptic segmentation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 108–126.
- [21] C.-F. Chen, Q. Fan, and R. Panda, “Crossvit: Cross-attention multi-scale vision transformer for image classification,” *arXiv preprint arXiv:2103.14899*, 2021.
- [22] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, “Multiscale vision transformers,” *arXiv preprint arXiv:2104.11227*, 2021.
- [23] J. Wang, Z. Wu, J. Chen, and Y.-G. Jiang, “M2tr: Multi-modal multi-scale transformers for deepfake detection,” *arXiv preprint arXiv:2104.09770*, 2021.
- [24] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, “Doubleu-net: A deep convolutional neural network for medical image segmentation,” in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2020, pp. 558–564.
- [25] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv preprint arXiv:1706.05587*, 2017.
- [26] N. K. Tomar, D. Jha, M. A. Riegler, H. D. Johansen, D. Johansen, J. Rittscher, P. Halvorsen, and S. Ali, “Fanet: A feedback attention network for improved biomedical image segmentation,” *arXiv preprint arXiv:2103.17235*, 2021.
- [27] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” *arXiv preprint arXiv:2012.12877*, 2020.
- [28] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in *European conference on computer vision*. Springer, 2016, pp. 354–370.
- [29] S. Nah, T. Hyun Kim, and K. Mu Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3883–3891.
- [30] C.-F. Chen, Q. Fan, N. Mallinar, T. Sercu, and R. Feris, “Big-little net: An efficient multi-scale feature representation for visual and speech recognition,” *arXiv preprint arXiv:1807.03848*, 2018.
- [31] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, “Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5386–5395.
- [32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
- [33] D. Jha, P. H. Smedsrød, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, “Kvasir-seg: A segmented polyp dataset,” in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.
- [34] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.
- [35] N. Tajbakhsh, S. R. Gurudu, and J. Liang, “Automated polyp detection in colonoscopy videos using shape and context information,” *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 630–644, 2015.
- [36] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzał, and A. Courville, “A benchmark for endoluminal scene segmentation of colonoscopy images,” *Journal of healthcare engineering*, vol. 2017, 2017.
- [37] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, “Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer,” *International journal of computer assisted radiology and surgery*, vol. 9, no. 2, pp. 283–293, 2014.
- [38] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, “Pranet: Parallel reverse attention network for polyp segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 263–273.

- [39] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, “Real-time polyp detection, localization and segmentation in colonoscopy using deep learning,” *Ieee Access*, vol. 9, pp. 40496–40510, 2021.
- [40] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic),” *arXiv preprint arXiv:1902.03368*, 2019.
- [41] P. Tschandl, C. Rosendahl, and H. Kittler, “The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.
- [42] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin *et al.*, “Nucleus segmentation across imaging experiments: the 2018 data science bowl,” *Nature methods*, vol. 16, no. 12, pp. 1247–1253, 2019.
- [43] D. Jha, P. H. Smedsrød, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, “Resunet++: An advanced architecture for medical image segmentation,” in *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2019, pp. 225–2255.
- [44] 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.
- [45] 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.
- [46] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, “Deep high-resolution representation learning for visual recognition,” *IEEE transactions on pattern analysis and machine intelligence*, 2020.
- [47] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.
- [48] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, “Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps,” *arXiv preprint arXiv:2101.07172*, 2021.
- [49] Y. Fang, C. Chen, Y. Yuan, and K.-y. Tong, “Selective feature aggregation network with area-boundary constraints for polyp segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 302–310.
- [50] Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual u-net,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.
- [51] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, “Bi-directional convlstm u-net with densely connected convolutions,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
- [52] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.