Stepwise Feature Fusion: Local Guides Global

Jinfeng Wang[1,2], Qiming Huang[1], Feilong Tang[1], Jia Meng[1], Jionglong Su[1(✉)], and Sifan Song[1,2(✉)]

[1] Xi'an Jiaotong-Liverpool University, Suzhou, China
[2] University of Liverpool, Liverpool, UK
Jionglong.Su@xjtlu.edu.cn, Sifan.Song19@student.xjtlu.edu.cn
https://github.com/Qiming-Huang/ssformer

Abstract. Colonoscopy, currently the most efficient and recognized colon polyp detection technology, is necessary for early screening and prevention of colorectal cancer. However, due to the varying size and complex morphological features of colonic polyps as well as the indistinct boundary between polyps and mucosa, accurate segmentation of polyps is still challenging. Deep learning has become popular for accurate polyp segmentation tasks with excellent results. However, due to the structure of polyps image and the varying shapes of polyps, it is easy for existing deep learning models to overfit the current dataset. As a result, the model may not process unseen colonoscopy data. To address this, we propose a new state-of-the-art model for medical image segmentation, the SSFormer, which uses a pyramid Transformer encoder to improve the generalization ability of models. Specifically, our proposed Progressive Locality Decoder can be adapted to the pyramid Transformer backbone to emphasize local features and restrict attention dispersion. The SSFormer achieves state-of-the-art performance in both learning and generalization assessment.

Keywords: Polyp segmentation · Deep learning · Generalization

1 Introduction

Colorectal cancer (CRC) is common cancer whose cancer risk may be reduced through early screening and removal of colon polyps [6,9]. However, accurate polyp segmentation is still a challenge due to the variable size and shape of polyps, as well as the indistinct boundaries between polyps and mucosa [6]. An accurate segmentation algorithm based on deep learning can effectively improve the accuracy and efficiency of polyp segmentation. Many image segmentation models based on the Convolutional Neural Networks (CNN) recently achieved excellent learning ability in several polyp segmentation benchmarks. [6,9,11,16,

J. Wang, Q. Huang and F. Tang—Contributed equally.

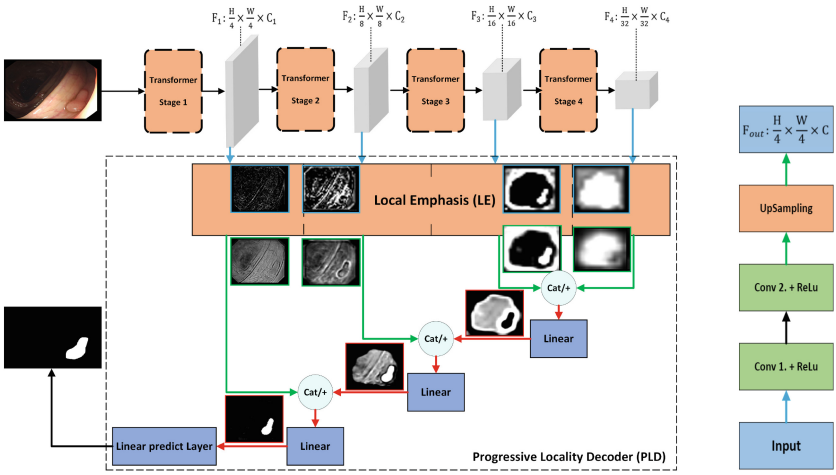
L. Wang et al. (Eds.): MICCAI 2022, LNCS 13433, pp. 110–120, 2022.
https://doi.org/10.1007/978-3-031-16437-8_11

25] However, due to the top-down modeling method of the CNN model and the variability in the morphology of polyps but relatively simple structure of the polyps image, this model lacks generalization ability and is difficult to process unseen datasets. To improve the generalization ability of the deep learning model, we shall incorporate the Transformer architecture into the polyp segmentation task.

The Transformer [18] was initially proposed as a bottom-up model architecture in the natural language processing (NLP) community. Dosovitskiy *et al.*. proposed the Vision Transformer (ViT) [5] that achieved superior performance in image classification tasks. The Transformer is different from CNN which the weight parameters are trained in the kernel to extract and mix the features among elements in the receptive field. In contrast, the Transformer obtains similarities of all patch pairs through the dot product between the patch vectors to adaptively extract and mix features between all patches. This enables the Transformer to have an efficient global receptive field and reduces the inductive bias of the model. As a result, the Transformer has a more robust generalization ability than CNN and Multilayer Perceptron-like structures [12]. However, the low inductive bias and powerful global receptive field make it difficult for the Transformer model to capture task-specific critical local details adequately. In addition, with the deepening of the Transformer model, the global features are continuously mixed and converged [24], resulting in attention dispersion. These make it difficult for the Transformer model to accurately predict detailed information in the dense prediction task of semantic segmentation.

In order to achieve high generalization and accurate polyp automatic segmentation, a novel state-of-the-art (SOTA) medical image segmentation model, SSFormer, is proposed which uses a pyramid Transformer encoder [10,19,20,22] for excellent generalization and multi-scale feature processing capabilities. In our model, the Progressive Locality Decoder (PLD), based on a multi-stage feature aggregation structure, functions as the decoder. The multi-stage feature aggregation structure can enable features of different depths and expressive powers to guide each other, which we believe can address the problems of attention dispersion and underestimation of local features to improve the detail processing ability. Segformer [22] optimized the encoder of the pyramid structure of PVT [19] and proposed a multi-stage feature aggregation decoder, which predicts features of different scales and depths separately through simple upsampling and then parallel fusion. SETR [23] uses the traditional Transformer as the encoder and proposes an MLA decoder with a multi-stage feature aggregation structure. Their excellent performances demonstrate that the decoding method of multi-stage feature aggregation is beneficial to improving the performance of Transformer in dense prediction tasks. Our proposed PLD adopts a stepwise adaptive method to emphasise local features and integrate them into global features, making the fusion of features more efficient.

The main contributions of this paper are: 1) We introduce the pyramid Transformer architecture into the polyp segmentation task to increase the generalization ability of the neural network; 2) We propose a new decoder PLD suitable for



(a) Overview of SSFormer (b) LE

Fig. 1. (a) The overview of SSFormer; (b) the structure of Local Emphasis module. In this figure, The lines with arrows and the feature maps next to them represent unemphasized features, local emphasized features, and fused features from top to bottom along the feature stream direction, respectively. The remainder of the PLD in Figure (a), excluding the Local Emphasis (LE), is the Stepwise Feature Aggregation (SFA). Feature fusion units can use concatenation (Cat) or addition (+) operations.

Transformer feature pyramids, which can smooth and effectively emphasise the local features in the Transformer to improve the detailed information processing ability of the neural network; 3) Our proposed SSFormer improves the SOTA performances of the ETIS benchmark, CVC-ClinicDB benchmark, and Kvasir benchmark by about 3%, 1.8%, and 1%, respectively. In addition, SSFormer has achieved state-of-the-art and superior performance in 2018 Data Science Bowl and ISIC-2018 benchmarks.

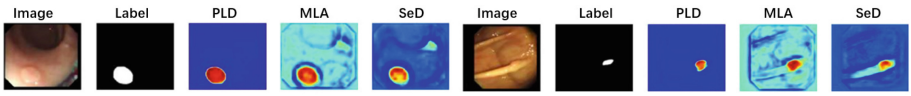
2 Methodology

2.1 Transformer Encoder

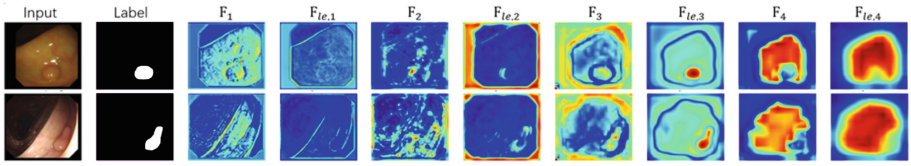
In order for our model to have enough generalization ability and multi-scale feature processing ability to carry out polyp segmentation, we use the Transformer based on the pyramid structure instead of CNN as the encoder. To this end, we adopt the encoder design of PVTv2 [20] and Segformer to construct the encoder. They both use the convolution operation to replace the PE operation of the traditional Transformer for consistency of spatial information, excellent performance and stability.

2.2 Aggregate Local and Global Features Stepwise (PLD)

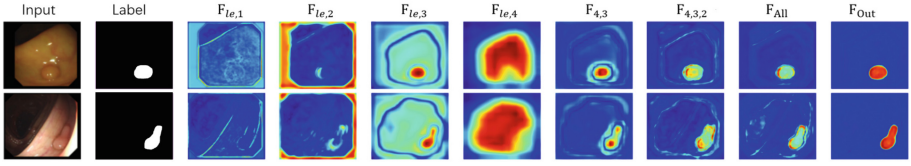
Experiments [13,23] have demonstrated that the sufficiency of local features obtained in the shallow part of the Transformer directly affects the performance of the model. However, we believe that the existing Transformer model lacks local and detailed information processing ability to focus on critical detailed features (such as contour, veins and texture). As a result, this makes it difficult for the model to locate the more decisive local feature distribution (mucosa can be considered a distribution composed of local features such as unique veins and textures). We propose a novel multi-stage feature aggregation decoder PLD for feature pyramids to address this issue. Figure 1(a) shows that the PLD consists of the Local Emphasis (LE) module and the Stepwise Feature Aggregation (SFA) module. The experimental section compares PLD with other existing decoders with various encoders that can generate feature pyramids. We compare the attention distribution before the final prediction of several typical multi-stage feature aggregation decoders for Transformers. As demonstrated in Fig. 2(a), after PLD fuses multi-stage features, the prediction head can accurately focus on critical targets. In addition, our PLD can be used for other Pyramid Transformer encoders and can improve the model's accuracy. There is a further demonstration in Sect. 3.3.



(a) Attention heatmap for different decoder(SeD is Segformer's Decoder)



(b) Attention Heatmap for the LE module



(c) Attention Heatmap for the SFA

Fig. 2. Attention heatmap of feature flow through the PLD process. Figure (a) shows that the LE module successfully focuses the model's attention on critical details. Figure (b) shows that the SFA structure effectively constrains the model's chaotic attention stepwise to fine critical regions.

Local Emphasis. In the Transformer, each patch in the image will mix the information of all other patches, even if their correlation is not high. After a large number of self-attention operations, the feature streams will converge, further exacerbating the attention dispersion or attention collapse [24]. Furthermore, we argue that the attention matrix in the self-attention mechanism can be viewed as a global non-preset convolution kernel. We designed the LE module using the local receptive field of the convolution kernel to increase the macro weights of the patches around the query patch to refocus attention on neighboring features thus reducing attention dispersion. In Fig. 1(b), the module consists of the convolution operators, activation functions, and a bilinear upsampling layer. We utilize the fixed receptive field of the convolution operator to mix the features of the adjacent patches of each patch, thereby increasing the associated weights of the adjacent patches to the center patch, thus emphasizing the local features of each patch. Since the feature types of the feature streams from different depths are different, we do not share the convolution weights for the feature streams at different levels in the feature pyramid. The formula for strengthening local features is as follows:

$$F_{le,i} = ReLU(Conv_i(C, C)(ReLU(Conv_i(C_i, C)(F_i)))), \tag{1}$$

where $F_{le,i}$ refer to the local emphasized feature from stage i, $Conv_i(C_{in}, C_{out})$ and $Linear(C_{in}, C_{out})$ refer to a convolutional and linear layer with input channel C_{in} and output channel C_{out}. From the feature map given in Fig. 1(a), it can be seen that the LE can effectively clean up cluttered noises and emphasize critical local features. In Fig. 2(b), after the feature stream passes through LE, the disordered attention is re-condensed along with critical details such as contours and boundaries.

Stepwise Feature Aggregation (SFA). Experiments [13] have demonstrated that the amount of information interacted through residual connections [7] in the Transformer is more significant than that of the CNN model. This phenomenon can be understood as the weak correlation between the features of different depths in the Transformer, requiring a lot of information interaction for the layers of different depths to guide each other. As such, we believe that direct parallel aggregation of features of different stages with significant differences in depth in Transformer may generate an information gap.

In order for the feature aggregation to be as smooth as possible, the SFA progressively fuses the features of different levels in the feature pyramid from the top to bottom. From the perspective of the change of feature streams, it can be considered that the local features of the shallower layers are progressively fused into the global features of the deeper layer. This feature fusion method can reduce the information gap between the fused high-dimensional and low-dimensional features. As given in Fig. 2(c), local features gradually guide the attention of the model to critical regions in the SFA. In Fig. 1(a), the SFA consists of feature fusion units, linear fusion layers, and a linear prediction layer. The feature map of the fused structure in Fig. 1(a) (image with red border) shows that the SFA

effectively incorporates local features into high-dimensional features and guides the feature stream into critical regions.

$$F_{i-1,i} = \begin{cases} Linear(2C, C)(Concat(F_{i-1}, F_i)), \\ OR, \\ Linear(C, C)(Add(F_{i-1}, F_i)), \end{cases} \tag{2}$$

Since the feature stream has the same shape after passing through the LE module, we can use concatenation or addition operation in the feature fusion unit as Eq. 2. In Table 4, we see that both perform equally well. Concatenation is the default in SSFormer.

2.3 Stepwise Segmentation Transformer

Based on the different encoder scales, we propose the SSFormer-S (Standard) and the SSFormer-L (Large) model. They achieve SOTA and competitive performance in several polyp segmentation benchmarks. Details are given in the experimental section. Moreover, SSFormer also achieved SOTA and competitive performance in ISIC-2018 and 2018 DATA Science Bowl.

3 Experiments

3.1 Experimental Setup

Dataset and Evaluation Matrix. Since the colon polyp segmentation task requires the model to have both accurate prediction and generalization capabilities, the performance of model on experimental and unseen benchmark datasets needs to be assessed separately. Therefore, following the experimental scheme of MSRF-Net [16], we train and test SSFormer on the Kavsir-SEG [8] and CVC-ClinicDB [1] benchmark datasets, respectively, to assess the accurate prediction and learning ability of models in the Kavsir-SEG and CVC-ClinicDB test set, respectively. In order to assess the generalization ability of SSFormer, we tested the model trained in Kavsir-SEG on CVC-ClinicDB and vice versa.

We refer to the experimental scheme of PraNet [6] and UACANet [9] that randomly extract 1450 images from the Kavsir and CVC-ClinicDB benchmark datasets to construct a training set (For fairness evaluation, we used the same training set as UACANet and PraNet), then test the model trained in this training set on the CVC-ColonDB [2] and ETIS [15] benchmark datasets. This test can demonstrate our model's accurate prediction and generalization ability in unseen datasets. Due to the variety of types and sizes of polyps in ETIS, it is the most challenging benchmark. The ISIC-2018 [4,17] and 2018 Data Science Bowl [3] benchmark datasets were also used in additional experiments. To unify the performance measures of the above two schemes, we only use mean Dice and mean IoU as evaluation matrices in our experimentation.

Implementation Details. We implement our model in PyTorch, which an NVIDIA TESLA A100 GPU accelerates. The AdamW optimizer is used with an initial learning rate of 0.0001, a decay rate of 0.1, and a decay period of 40 epochs. The training period is 200 epochs. Our loss function is the combined loss of Dice loss and BCE loss. During training, we resize the image to 352×352. We employ random flipping, scaling, rotation, and random dilation and erosion as data augmentation operations.

Table 1. The performance of the SOTA methods was trained and tested on the same benchmark dataset, used to assess learning ability, the scores in the table refer to [16]

Dataset	CVC-ClinicDB		Kvasir-SEG		ISIC-2018		2018 Data-Sci Bowl	
Methods	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net	0.9145	0.8654	0.8629	0.8176	0.8554	0.7847	0.9080	0.8314
U-Net++	0.8453	0.7559	0.7475	0.6313	0.8094	0.7288	0.7705	0.3010
Deeplabv3+	0.8897	0.8706	0.8965	0.8575	0.8772	0.8128	0.8857	0.8367
MSRF-Net	0.9420	**0.9043**	0.9217	**0.8914**	0.8824	0.8373	0.9224	0.8534
SSFormer-S	0.9268	0.8759	0.9261	0.8743	0.9195	0.8615	**0.9254**	**0.8652**
SSFormer-L	**0.9447**	0.8995	**0.9357**	0.8905	**0.9242**	**0.8675**	0.9230	0.8614

3.2 Results

Learning Ability. We split the CVC-ClinicDB and Kvasir benchmark datasets into 80% training set, 10% evaluation set and 10% test set according to the first scheme mentioned in Sect. 3.1. Table 1 demonstrates that our model improves the SOTA result by about 1.8% on the CVC-ClinicDB benchmark and about 1% on the Kvasir benchmark. These performances demonstrate the superior accurate prediction and learning abilities of SSFormer.

Furthermore, to assess the performance of SSFormer on other medical segmentation benchmarks, we conduct additional experiments on the ISIC-2018 and 2018 Data Science Bowl benchmark datasets. The results in Table 1 reveal that our model achieves the SOTA and excellent performance on two benchmarks, 2018 Data Science and ISIC-2018, respectively.

Generalization Ability. We test the SSFormer trained on the CVC-ClinicDB and Kvasir datasets on the Kvasir and CVC-ClinicDB benchmarks, respectively. As mentioned in Sect. 3.1, this test result can reflect the generalization ability of our model. In Table 2, our model achieves outstanding performance using this testing scheme. In addition, to further assess the generalization ability of SSFormer, we refer to the experimental scheme of PraNet, use the training set constructed from part of the Kvasir and CVC-ClinicDB datasets for training,

Table 2. Generalization Test 1

Train set	CVC-ClinicDB		Kvasir-SEG	
Test set	Kvasir-SEG		CVC-ClinicDB	
Methods	mDice	mIoU	mDice	mIoU
U-Net	0.6222	0.4588	0.7172	0.6133
U-Net++	0.5926	0.4564	0.4265	0.3345
Deeplabv3+	0.6746	0.5327	0.6509	0.5385
MSRF-Net	0.7575	0.6337	0.7921	0.6498
SSFormer-S	0.7790	0.6977	0.7966	0.7229
SSFormer-L	**0.8270**	**0.7348**	**0.8339**	**0.7573**

Table 3. Generalization Test 2

Train set	Kvasir & CVC-ClinicDB			
Test set	CVC-ColonDB		ETIS	
Method	mDice	mIoU	mDic	mIoU
UACANet-S	0.783	0.704	0.694	0.615
UACANet-L	0.751	0.678	0.766	0.689
CaraNet	0.773	0.689	0.747	0.672
PraNet	0.712	0.640	0.628	0.567
SSformer-S	0.772	0.697	0.767	0.698
SSformer-L	**0.802**	**0.721**	**0.796**	**0.720**

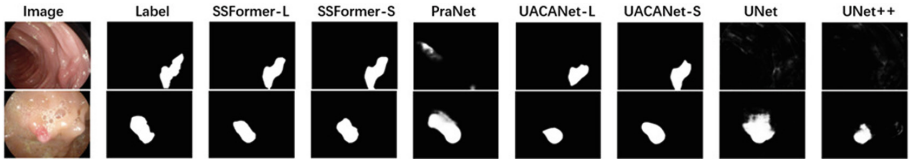


Fig. 3. Predicted results of different methods

and test the model on the CVC-ColonDB and ETIS benchmarks. The results in Table 3 demonstrate that our model significantly improves the SOTA performance (3%) in the most challenging ETIS and achieves superior performance in CVC-ColonDB. Figure 3 gives the prediction accuracy of our model on the ETIS benchmark. These results can prove that SSFormer has robust generalization and accurate prediction abilities. (The scores in Table 2 and Table 3 are obtained from [9,14,16,25]).

Table 4. Different encoder and decoder combinations performance with the same hyperparameter settings. The performance of different encoder and decoder combinations. The score is the performance of the model on the (CVC-ClinicDB, Kvasir) dataset group. (SeD is Segformer's Decoder, MiT is the Segformer's Encoder, and the CvT is proposed in [21])

Encoder\Decoder	MLA [23]	SeD [22]	PLD-Cat	PLD-Add
CvT [21]	0.898, 0.912	0.820, 0.889	0.912, 0.923	-
PvT [19]	0.809, 0.799	0.588, 0.618	0.828, 0.801	-
MiT [22]	0.907, 0.893	0.911, 0.903	0.916, 0.925	0.923, 0.897

Table 5. The effect of PLD components on model performance. The score in the table follow (mDice, mIOU).

Decoder\Dataset	Kvasir-SEG	ISIC-2018	2018 Data-Science Bowl
Without PLD	0.869, 0.918	0.855, 0.894	0.835, 0.904
LE	0.877, 0.925	0.860, 0.909	0.850, 0.915
SFA	0.885, 0.930	0.863, 0.918	0.858, 0.920
LE+SFA	**0.891, 0.936**	**0.868, 0.924**	**0.861, 0.923**

3.3 Ablation Study

In Table 4, the PLD performs the best with the MiT. We believe that this is because the convolution operation inside MiT can maintain the consistency of the spatial information of the model. Furthermore, the experiments in Table 5 demonstrate the effectiveness of the PLD and its components.

4 Conclusions

In this research, we propose a novel deep learning model SSFormer, with robust generalization and learning ability. These are critical for polyp segmentation. Furthermore, we find that our model also demonstrates powerful learning ability in ISIC-2018 and 2018 Data Science Bowl benchmarks in additional experiments. We believe that the SSFormer has great potential to improve deep learning performance in other medical image segmentation tasks. Furthermore, experiments demonstrate that our proposed local feature emphasis module effectively constrains the attention dispersion of Transformers. Therefore, our research can be further used to optimize the Transformer backbone network for the general computer vision community and high generalizability medical applications.

Acknowledgments. This work was supported by the Key Program Special Fund in XJTLU (KSF-A-22).

References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. **43**, 99–111 (2015)
2. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recogn. **45**(9), 3166–3182 (2012)
3. Caicedo, J.C., et al.: Nucleus segmentation across imaging experiments: the 2018 data science bowl. Nat. Methods **16**(12), 1247–1253 (2019)
4. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172. IEEE (2018)

5. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Fan, D.-P., et al.: PraNet: parallel reverse attention network for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 263–273. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_26
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
9. Kim, T., Lee, H., Kim, D.: UACANet: uncertainty augmented context attention for polyp segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2167–2175 (2021)
10. Li, G., Xu, D., Cheng, X., Si, L., Zheng, C.: SimViT: exploring a simple vision transformer with sliding windows (2021)
11. Lou, A., Guan, S., Loew, M.: CaraNet: context axial reverse attention network for segmentation of small medical objects. arXiv preprint arXiv:2108.07368 (2021)
12. Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Shahbaz Khan, F., Yang, M.H.: Intriguing properties of vision transformers. In: Advances in Neural Information Processing Systems 34 (2021)
13. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? In: Advances in Neural Information Processing Systems 34 (2021)
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. **9**(2), 283–293 (2014)
16. Srivastava, A., et al.: MSRF-Net: a multi-scale residual fusion network for biomedical image segmentation. arXiv preprint arXiv:2105.07451 (2021)
17. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**(1), 1–9 (2018)
18. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)
19. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
20. Wang, W., et al.: PVT v2: improved baselines with pyramid vision transformer. Comput. Visual Media **8**, 415–424 (2022). https://doi.org/10.1007/s41095-022-0274-8
21. Wu, H., et al.: CvT: introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22–31 (2021)
22. Xie, E., et al.: SegFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems 34 (2021)

23. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
24. Zhou, D., et al.: DeepViT: towards deeper vision transformer. arXiv preprint arXiv:2103.11886 (2021)
25. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1