Diffusion Transformer U-Net for Medical Image Segmentation

G. Jignesh Chowdary(✉) and Zhaozheng Yin

Stony Brook University, Stony Brook, NY, USA
jigneshchowdary@gmail.com

Abstract. Diffusion model has shown its power on various generation tasks. When applying the diffusion model in medical image segmentation, there are a few roadblocks to remove: the semantic features required for the conditioning of the diffusion process are not well aligned with the noise embedding; and the U-Net backbone employed in these diffusion models is not sensitive to contextual information that is essential during the reverse diffusion process for accurate pixel-level segmentation. To overcome these limitations, we present a cross-attention module to enhance the conditioning from source images, and a transformer based U-Net with multi-sized windows for the extraction of various scales of contextual information. Evaluated on five benchmark datasets with different imaging modalities including Kvasir-Seg, CVC Clinic DB, ISIC 2017, ISIC 2018, and Refuge, our diffusion transformer U-Net achieves great generalization ability and outperforms all the state-of-the-art models on these datasets.

Keywords: Diffusion model · Transformer · U-Net · Medical Image Segmentation

1 Introduction

Deep Learning (DL) methods like Convolutional Neural Networks (CNN) and Vision-Transformers (ViT) have been applied to medical image segmentation [7,8,17] with good performance. However, these DL methods have some inherent limitations on their network architectures. For example, CNNs are capable of extracting local features but not direct global features, whereas ViTs employ a fixed window which limit their capability to extract fine contextual details that are necessary for accurate pixel-level segmentation.

Recently, Denoising Diffusion Probabilistic Model (DDPM) [9] shows great performance in various conditional and unconditional generation tasks, and it is also applied to medical image segmentation [23,24]. Despite of the success, there are a few shortcomings to overcome: (1) The semantic embedding extracted from the source image is not well aligned with the noise embedding in the diffusion

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_59.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 622–631, 2023.
https://doi.org/10.1007/978-3-031-43901-8_59

process, leading to poor conditioning and subpar performance; and (2) The U-Net backbone in these DDPM-based methods is not sensitive to various scales of contextual information during the reverse diffusion (denoising) process, observed in CNNs and ViTs as well.

Motivated by the underlined limitations, we propose a Diffusion Transformer U-Net, with the following contributions:

- A conditional diffusion model with forward and backward processes is proposed to train segmentation networks. In the backward denoising process, the feature embedding of a noise image is aligned with that of the conditional source image by a new cross-attention module. Then, it is denoised into a segmentation mask of the source image by the segmentation network.
- A transformer-based U-Net with multi-sized windows, named as *MT U-Net*, is designed to extract both pixel-level and global contextual features for achieving good segmentation performance.
- The MT U-Net trained by the diffusion model has a great generalization capability on various imaging modalities, and outperforms all the current state-of-the-art on five benchmark datasets including polyp segmentation from colonoscopy images [1,10], skin lesion segmentation from dermoscopy images [4,5], and optic-cup segmentation from retinal fundus images [14].

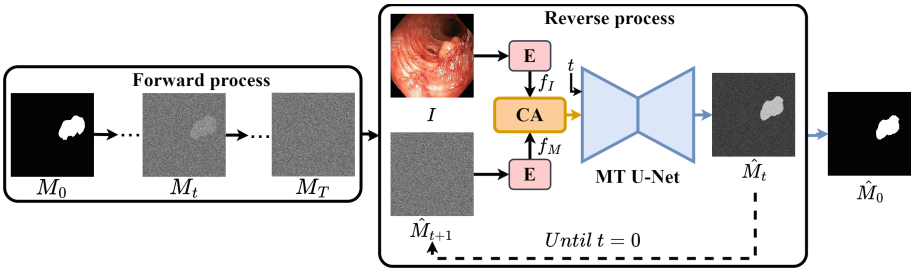

Fig. 1. Diffusion model with Cross-Attention (CA) to train the MT U-Net.

2 Method

2.1 Diffusion Model

The diffusion has two processes (Fig. 1): forward and reverse. In the forward process, the ground truth M_0 is transformed into noisy ground truth M_T by gradually adding Gaussian noise through T time steps. In the reverse process, first, the source image I and noise map $\hat{M}_{t+1}$ pass through an encoder E (two residual-inception blocks [18]) to obtain embedding $f_I \in R^{h \times w \times c_1}$ and $f_M \in R^{h \times w \times c_2}$ (subscripts I and M denote image and map), where h, w and c_1 (c_2) are the height, width and channels of the embeddings, respectively. Then, the two embeddings are aligned by a Cross-Attention (CA) module in the feature space. The aligned feature map is given as a noisy input to MT U-Net to recover

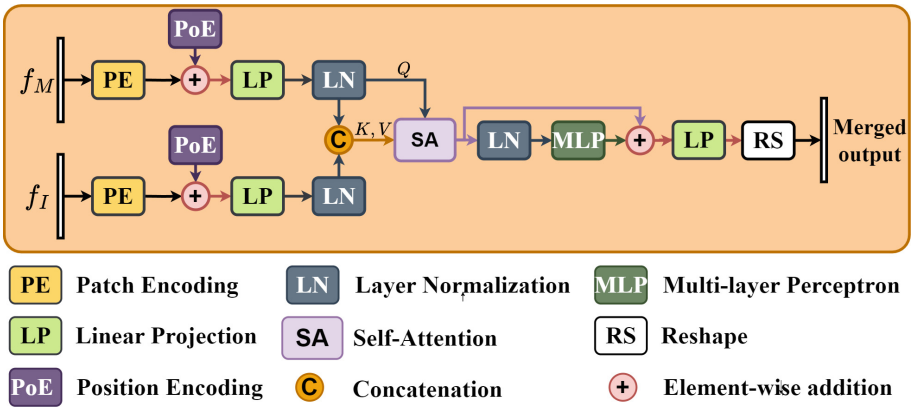

Fig. 2. Architecture of the Cross-Attention (CA) module.

$\hat{M}_t$. This reverse process iterates from $t = T - 1$ till $t = 0$ (i.e., the initial $\hat{M}_{t+1}$ when $t = T - 1$, $\hat{M}_T$, is set as M_T, and $\hat{M}_0$ is recovered eventually, which is expected to be identical to the ground truth M_0).

Figure 2 presents the architecture of our CA module, which is used to align f_M and f_I in order to improve the conditioning of the diffusion model. First, f_M and f_I are divided into patches and flattened to vectors by a Patch Encoding (PE) layer. Then, the position information of patches is obtained using a Position Encoding layer (PoE), and is added to the original patch embeddings for preserving their positional information. The dimensions of the two position-built-in patch embeddings are aligned using Linear Projection (LP) layers, and are normalized by a Layer Normalization (LN), denoting the output after the two LN's as $f_M^p \in R^d$ and $f_I^p \in R^d$ (d-dim feature vectors of $p$atches). Thirdly, we use a Self-Attention for efficient feature fusion:

$$SA = Softmax(\frac{QK^\top}{\sqrt{d}})V \tag{1}$$

where f_M^p is the query (Q), the concatenation of f_M^p and f_I^p is the key (K) and value (V). $\top$ denotes the transpose. Fourthly, following [20], we encode the output of L_{SA} by a Layer Normalization (LN) and a two-layered Multi-Layer Perceptron (MLP) for extracting more contextual information. An auxiliary connection (residual) is used to enhance the information propagation. Lastly, we apply a Reshape (RS) layer to reshape and assemble the patches into the same size as f_M.

2.2 Multi-sized Transformer U-Net (MT U-Net)

Figure 3(a) presents the architecture of our MT U-Net, with the encoding and decoding parts. The encoding part consists of a Patch Partitioning layer, a Linear Embedding layer, a PoE and four Encoder blocks. The Patch Partitioning layer splits the input into non-overlapping patches with a patch size of 2×2.

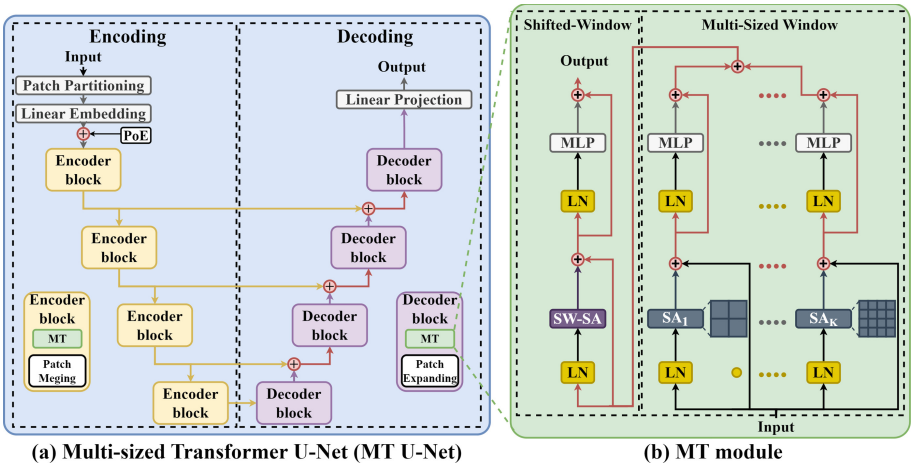

Fig. 3. Architecture of the proposed MT U-Net, and the MT module. The time step embedding is not presented in the figure for clarity.

These patches, along with the time embedding are flattened into a $D \times 1$ dimension linear embedding using a Linear Embedding layer. Then positional information obtained from the PoE is added to the linear embedding before passing through the four Encoder blocks. Each Encoder block consists of a Multi-sized Transformer (MT) module and a Patch Merging layer, except the last encoder block which only contains the MT module. The MT module extracts multi-scale contextual features (to be elaborated later), and the Patch Merging layer down-samples the feature maps. With the inspiration from U-Net [15], a skip connection is employed for using the multi-scale contextual information from the encoder to overcome the loss of spatial information during down-sampling. Similar to the Encoder block, each Decoder block consists of an MT module and a patch-expanding layer, except the first decoder block which only contains the MT module. The patch-expanding layer performs the up-sampling, and reshaping operation on feature maps. Finally, we employ a Linear Projection layer to obtain the pixel-level predictions.

The proposed Multi-sized Transformer (MT) module (Fig. 3(b)) is different from the conventional transformer [6]. The MT module consists of two parts: multi-sized window and shifted-window. The multi-sized window part extracts multi-scale contextual information, and the shifted-window part enriches the extracted information. The multi-sized window part has K parallel branches, with each branch consisting of a Layer Normalization (LN), multi-head Self-Attention (SA), auxiliary connection (residual), and a Multi-layer perceptron (MLP) with two layers followed by the GELU activation function. The window size used in the multi-head self-attention is varied to extract multi-scale contextual features. The output of these individual branches is combined, and is sent to the shifted-window part. The shifted-window part has a structure similar to

the individual branch in the multi-sized window, but it uses shifting windows in the self attention (SW-SA).

2.3 Training and Inference

During training, a source image and its segmentation ground truth map are given as input to the diffusion model. The diffusion model is trained using the noise prediction loss (L_{Noise}) [12] and cross-entropy loss (L_{CE}).

$$L_{TOTAL} = L_{Noise}(M_t, \hat{M}_t) + L_{CE}(M_t, \hat{M}_t) \tag{2}$$

During inference, a noise image sampled from the Gaussian distribution, along with the testing image, is given as the input to the reverse process.

3 Experimental Results

3.1 Datasets and Evaluation Metrics

To evaluate the effectiveness and generalization ability of the proposed method, different medical image segmentation tasks are tested, including: (1) **Polyp segmentation from colonoscopy images** (Kvasir-SEG (KSEG) [10], CVC-Clinic DB (CVC) [1]), (2) **Skin lesion segmentation from dermoscopy images** (ISIC 2017 (IS17') [5], ISIC 2018 (IS18') [4,19]), and (3) **Optic-cup segmentation from retinal fundus images** (REFUGE (REF) [14]). Dice Coefficient (DC) and Intersection over Union (IoU) are used as evaluation metrics.

3.2 Implementation Details

The number of branches in the MT module is set to 3 by cross-validation, with window sizes as 4, 8, and 16 respectively. The diffusion transformer U-Net is trained for 40,000 iterations using SGD optimizer with a momentum of 0.6, with a batch size of 16, and the learning rate is set to 0.0005. In the diffusion, we use a linear noise scheduler with $T = 1000$ steps. For fair comparisons with the recent diffusion-based segmentation models [23,24], during inference an average ensemble of 25 predictions is considered as the final prediction. All the experiments are conducted using a NVIDIA Tesla V-100 GPU with 32 GB RAM.

3.3 Performance Comparison

First, we quantitatively compare our method with several well-known U-Net and/or Transformer-related segmentation models, including U-Net [15], U-Net++ [26], Attention U-Net [13], Swin U-Net [2], Trans U-Net [3], and SegFormer [25]. With their source codes, these models are trained, and evaluated on the experiment datasets. For fair comparisons, all models use the same experimental protocol for each dataset. The quantitative results are shown in Table 1.

Table 1. Comparison with state-of-the-art methods related to U-Net and/or Transformers. '80 : 10 : 10' (data split on training:validation:testing) experimental protocol is employed on KSEG, CVC, IS18'; respective default splits are used on REF, and IS17'.

Metric	Datasets	Models						
		U-Net [15]	U-Net ++ [26]	Attention U-Net [13]	Swin U-Net [2]	Trans U-Net [3]	SegFormer [25]	Ours
DC	KSEG [10]	0.775	0.786	0.798	0.867	0.889	0.905	**0.946**
	CVC [1]	0.856	0.874	0.887	0.914	0.920	0.931	**0.954**
	IS18' [4,19]	0.813	0.833	0.844	0.869	0.904	0.914	**0.931**
	IS17' [5]	0.794	0.814	0.815	0.851	0.873	0.884	**0.935**
	REF [14]	0.769	0.779	0.792	0.815	0.821	0.843	**0.887**
IoU	KSEG [10]	0.714	0.725	0.752	0.854	0.863	0.874	**0.916**
	CVC [1]	0.805	0.821	0.845	0.874	0.889	0.905	**0.920**
	IS18' [4,19]	0.691	0.703	0.732	0.813	0.821	0.847	**0.879**
	IS17' [5]	0.659	0.663	0.682	0.721	0.745	0.768	**0.801**
	REF [14]	0.692	0.701	0.714	0.746	0.774	0.796	**0.815**

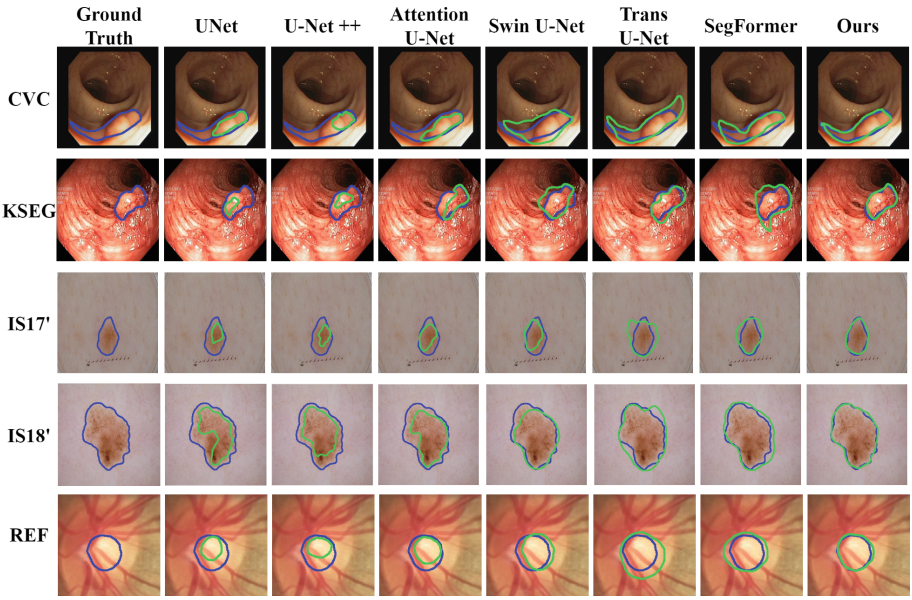

Fig. 4. Qualitative comparison with SOTA approaches on KSEG [10], CVC [1], IS18' [4,19], IS17' [5], and REF [14] datasets. The blue contours represent the ground truth, and the green contours represent the predicted results.

Our Diffusion Transformer U-Net outperforms all other U-Net or Transformer related models on the five datasets with various imaging modalities, validating its effectiveness and generalization capability.

Secondly, we qualitatively compare our Diffusion Transformer U-Net with other U-Net or Transformer related models. From the randomly sampled testing images in Fig. 4, we observe that the other models produce either over-segmented (e.g., Trans U-Net, SegFormer) or under-segmented results (e.g., U-

Table 2. Comparison with SOTA results. '-': No results reported. '*': number of images.

Dataset	Models	Publication, Year	Experimental protocol	DC	IoU
KSEG [10]	MSRF-Net [17]	IEEE JBHI, 2022	80:10:10	0.921	0.891
	Li-SegPNet [16]	IEEE TBE, 2022	80:10:10	0.905	0.828
	SSFormer [21]	MICCAI, 2022	80:10:10	0.935	0.890
	Ours		80:10:10	**0.946**	**0.916**
CVC [1]	MSRF-Net [17]	IEEE JBHI, 2022	80:10:10	0.942	0.904
	Li-SegPNet [16]	IEEE TBE, 2022	80:10:10	0.925	0.860
	SSFormer [21]	MICCAI, 2022	80:10:10	0.944	0.899
	Ours		80:10:10	**0.954**	**0.920**
IS18' [4,19]	MSRF-Net [17]	IEEE JBHI, 2022	80:10:10	0.882	0.837
	FAT-Net [22]	MedIA, 2022	1815*,259*,520*	0.890	0.820
	HiFormer [8]	WACV, 2023	1815*,259*,520*	0.910	-
	Ours		1815*,259*,520*	**0.924**	**0.843**
			80:10:10	**0.931**	**0.879**
IS17' [5]	FAT-Net [22]	MedIA, 2022	Default split	0.850	0.765
	HiFormer [8]	WACV, 2023	Default split	0.925	-
	ConTrans [11]	MICCAI, 2022	Default split	0.875	-
	Ours		Default split	**0.935**	**0.801**
REF [14]	MedSegDiff [23]	Arxiv, 2022	Default Split	0.863	0.782
	MedSegDiff-V2 [24]	Arxiv, 2023	Default Split	0.859	0.796
	Ours		Default Split	**0.887**	**0.815**

Net, U-Net++, Attention U-Net, Swin U-Net), and our segmentation masks are closest to the ground truth, demonstrating the effectiveness of our method.

Lastly, we compare our Diffusion Transformer U-Net with all the latest best models on the five datasets, as summarized in Table 2. The results from cited methods are copied from their papers directly, except for MedSegDiff, and MedSegDiff-V2. These two approaches are re-trained, and evaluated on the REF dataset. Note, since some methods use different experiment protocols on the IS18' dataset. For fair comparisons, we train/cross-validate/test our method using two different protocols, and compare ours with other methods with the same protocol. As shown in Table 2, our method consistently outperforms all the current best models on these five datasets, which again verifies its effectiveness and superiority.

3.4 Ablation Study

We perform a set of ablation studies to evaluate the contribution of each module in our Diffusion Transformer U-Net, as shown in Table 3:

– The original U-Net [15] is used as a baseline (row 1 in Table 3).

Table 3. Ablation on KSEG [10], CVC [1], IS18' [4,19], IS17' [5], and REF [14].

U-Net	Diff	CA	Vanila Trans	MT	IoU					DC				
					KSEG	CVC	IS17'	IS18'	REF	KSEG	CVC	IS17'	IS18'	REF
✓	✗	✗	✗	✗	0.714	0.805	0.659	0.691	0.692	0.775	0.856	0.794	0.813	0.769
✓	✓	✗	✗	✗	0.843	0.842	0.734	0.801	0.732	0.873	0.879	0.853	0.885	0.815
✓	✓	✓	✗	✗	0.875	0.875	0.763	0.841	0.763	0.905	0.932	0.889	0.901	0.831
✓	✓	✓	✓	✗	0.889	0.894	0.784	0.863	0.782	0.921	0.939	0.914	0.913	0.847
✓	✓	✓	✗	✓	**0.916**	**0.920**	**0.801**	**0.879**	**0.815**	**0.946**	**0.954**	**0.935**	**0.931**	**0.887**

- We replace the Cross-Attention (CA) in our diffusion model by a simple concatenation operation, and apply this simplified diffusion model to train the U-Net. Even this simplified diffusion model (row 2) can boost the U-Net performance in row 1, showing the effectiveness of diffusion.
- Using our diffusion model with the CA module (row 3), the performance is further improved, compared to the basic concatenation operation (row 2), which validates the contribution of the CA model for aligning feature embeddings during the denoising process of the diffusion model.
- Using our diffusion model with the CA module, we add the basic transformer units [6] without the multi-sized window into the U-Net (row 4). This also increases the segmentation performance, compared to row 3, which demonstrates that transformers can help the U-Net on segmentation.
- Based on the model from row 4, we add multi-sized windows into the transformer (i.e., our Diffusion Transformer U-Net, row 5). This gives the highest performance, compared to other configurations in the ablation studies.

4 Conclusion

A Diffusion Transformer U-Net is proposed for medical image segmentation. Instead of a standard U-Net in the diffusion model, we propose a transformer based U-Net with multi-sized windows for enhancing the contextual information extraction and reconstruction. We also design a cross-attention module to align feature embeddings, providing a better conditioning from the source image to the diffusion model. The evaluation on various datasets of different modalities shows the effectiveness and generalization ability of the proposed method.

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. Cao, H., et al.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: ECCV 2022, Part III, pp. 205–218. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_9

3. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Codella, N., 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)
5. 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)
6. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Gu, R., et al.: Ca-net: comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging **40**(2), 699–711 (2020)
8. Heidari, M., et al.: Hiformer: hierarchical multi-scale representations using transformers for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6202–6212 (2023)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
10. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., Cheng, W.-H., Kim, J., Chu, W.-T., Cui, P., Choi, J.-W., Hu, M.-C., De Neve, W. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
11. Lin, A., Xu, J., Li, J., Lu, G.: Contrans: improving transformer with convolutional attention for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V. pp. 297–307. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_29
12. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR (2021)
13. Oktay, O., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
14. Orlando, J.I., et al.: Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. **59**, 101570 (2020)
15. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Sharma, P., Gautam, A., Maji, P., Pachori, R.B., Balabantaray, B.K.: Li-segpnet: encoder-decoder mode lightweight segmentation network for colorectal polyps analysis. IEEE Trans. Biomed. Eng. (2022)
17. Srivastava, A., et al.: Msrf-net: a multi-scale residual fusion network for biomedical image segmentation. IEEE J. Biomed. Health Inform. **26**(5), 2252–2263 (2021)
18. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
19. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)

20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)
21. Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., Song, S.: Stepwise feature fusion: Local guides global. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III. pp. 110–120. Springer (2022)
22. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: Fat-net: feature adaptive transformers for automated skin lesion segmentation. Med. Image Anal. **76**, 102327 (2022)
23. Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y.: Medsegdiff: medical image segmentation with diffusion probabilistic model. arXiv preprint arXiv:2211.00611 (2022)
24. Wu, J., Fu, R., Fang, H., Zhang, Y., Xu, Y.: Medsegdiff-v2: diffusion based medical image segmentation with transformer. arXiv preprint arXiv:2301.11798 (2023)
25. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: simple and efficient design for semantic segmentation with transformers. Adv. Neural. Inf. Process. Syst. **34**, 12077–12090 (2021)
26. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging **39**(6), 1856–1867 (2019)