ConvFormer: Plug-and-Play CNN-Style Transformers for Improving Medical Image Segmentation

Xian Lin[1], Zengqiang Yan[1(✉)], Xianbo Deng[2], Chuansheng Zheng[2], and Li Yu[1]

[1] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China
{xianlin,z_yan,hustlyu}@hust.edu.cn
[2] Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology,Wuhan, China
dengxianbo@hotmail.com, cszheng@hust.edu.cn

Abstract. Transformers have been extensively studied in medical image segmentation to build pairwise long-range dependence. Yet, relatively limited well-annotated medical image data makes transformers struggle to extract diverse global features, resulting in attention collapse where attention maps become similar or even identical. Comparatively, convolutional neural networks (CNNs) have better convergence properties on small-scale training data but suffer from limited receptive fields. Existing works are dedicated to exploring the combinations of CNN and transformers while ignoring attention collapse, leaving the potential of transformers under-explored. In this paper, we propose to build CNN-style Transformers (ConvFormer) to promote better attention convergence and thus better segmentation performance. Specifically, ConvFormer consists of pooling, CNN-style self-attention (CSA), and convolutional feed-forward network (CFFN) corresponding to tokenization, self-attention, and feed-forward network in vanilla vision transformers. In contrast to positional embedding and tokenization, ConvFormer adopts 2D convolution and max-pooling for both position information preservation and feature size reduction. In this way, CSA takes 2D feature maps as inputs and establishes long-range dependency by constructing self-attention matrices as convolution kernels with adaptive sizes. Following CSA, 2D convolution is utilized for feature refinement through CFFN. Experimental results on multiple datasets demonstrate the effectiveness of ConvFormer working as a plug-and-play module for consistent performance improvement of transformer-based frameworks. Code is available at https://github.com/xianlin7/ConvFormer.

Keywords: CNN-Style Transformers · Attention Collapse · Adaptive Self-Attention · Medical Image Segmentation

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_61.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 642–651, 2023.
https://doi.org/10.1007/978-3-031-43901-8_61

1 Introduction

Benefiting from the prominent ability to model long-range dependency, transformers have become the de-facto standard for natural language processing [1]. Compared with convolutional neural networks (CNNs), which encourage locality, weight sharing, and translation equivariance, transformers build global dependency through self-attention layers, bringing more possibilities for feature exaction and breaking the performance ceiling of CNNs in return [2–6].

Inspired by this, transformers are introduced into medical image segmentation and arouse wide concerns [7–11]. In vision transformers, each medical image is first split into a series of patches and then projected into a 1D sequence of patch embeddings [4]. Through building pairwise interaction among patches/tokens, transformers are supposed to aggregate global information for robust feature exaction. However, learning well-convergence global dependency in transformers is highly data-intensive, making transformers less effective given relatively limited medical imaging data.

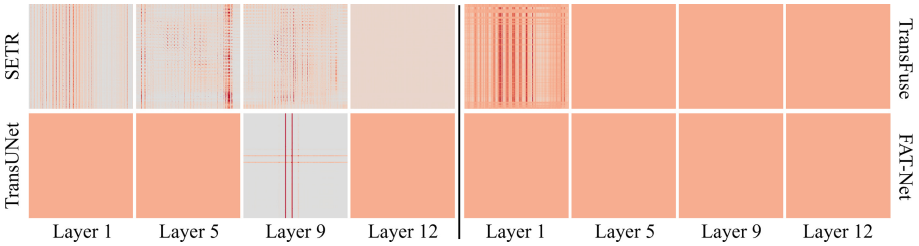

Fig. 1. Visualization of attention maps from the selected layers of the first head in different transformer frameworks. The darker the color, the closer the dependency.

To figure out how transformers work in medical image segmentation, we trained four state-of-the-art transformer-based models [5,12–14] on the ACDC dataset and visualized the learned self-attention matrices across different layers as illustrated in Fig. 1. For all approaches, the attention matrices tend to become uniform among patches (*i.e.*, attention collapse [15]), especially in deeper layers. Attention collapse is more noticeable, especially in CNN-Transformer hybrid approaches (*i.e.*, TransUNet, TransFuse, and FAT-Net). On the one hand, insufficient training data would make transformers learn sub-optimal long-range dependency. On the other hand, directly combining CNNs with transformers would make the network biased to the learning of CNNs, as the convergence of CNNs is more achievable compared to transformers, especially on small-scale training data. Therefore, how to address attention collapse and improve the convergence of transformers is crucial for performance improvement.

In this work, we propose a plug-and-play module named ConvFormer to address attention collapse by constructing a kernel-scalable CNN-style transformer. In ConvFormer, 2D images can directly build sufficient long-range dependency without being split into 1D sequences. Specifically, corresponding to tok-

enization, self-attention, and feed-forward network in vanilla vision transformers, ConvFormer consists of pooling, CNN-style self-attention (CSA), and convolutional feed-forward network (CFFN) respectively. For an input image/feature map, its resolution is first reduced by applying convolution and max-pooling alternately. Then, CSA builds appropriate dependency for each pixel by adaptively generating a scalable convolutional, being smaller to include locality or being larger for long-range global interaction. Finally, CFFN refines the features of each pixel by applying continuous convolutions. Extensive experiments on three datasets across five state-of-the-art transformer-based methods validate the effectiveness of ConvFormer, outperforming existing solutions to attention collapse.

2 Related Work

Recent transformer-based approaches for medical image analysis mainly focus on introducing transformers for robust features exaction in the encoder, cross-scale feature interactive in skip connection, and multifarious feature fusion in the decoder [16–19]. The study about addressing attention collapse for transformers in medical imaging is under-explored. Even in natural image processing, attention collapse, usually existing in the very deep layers of deep transformer-based models, has not been fully studied. Specifically, Zhou *et al.* [15] developed Re-attention to re-generate self-attention matrices aiming at increasing their diversity on different layers. Zhou *et al.* [20] projected self-attention matrices into a high-dimensional space and applied convolutions to promote the locality and diversity of self-attention matrices. Touvron *et al.* [21] proposed to re-weight the channels of the outputs from the self-attention module and the feed-forward module to facilitate the convergence of transformers.

3 Method

The comparison between the vision transformer (ViT) and ConvFormer is illustrated in Fig. 2. The greatest difference is that our ConvFormer is conducted on 2D inputs while ViT is applied to 1D sequences. Specifically, the pooling module is utilized to replace tokenization in ViT, which well preserves locality and positional information without extra positional embeddings. The CNN-style self-attention (CSA) module, *i.e.* the core of ConvFormer, is developed to replace the self-attention (SA) module in ViT to build long-range dependency by constructing self-attention matrices in a similar way like convolutions with adaptive and scalable kernels. The convolutional feed-forward network (CFFN) is developed to refine the features for each pixel corresponding to the feed-forward network (FFN) in ViT. No upsampling procedure is adopted to resize the output of ConvFormer back to the input size as the pooling module can match the output size by adjusting the maxpooling times. It should be noticed that ConvFormer is realized based on convolutions, which eliminates the training tension between CNNs and transformers as analyzed in Sect. 1. Each module of ConvFormer is described in the following.

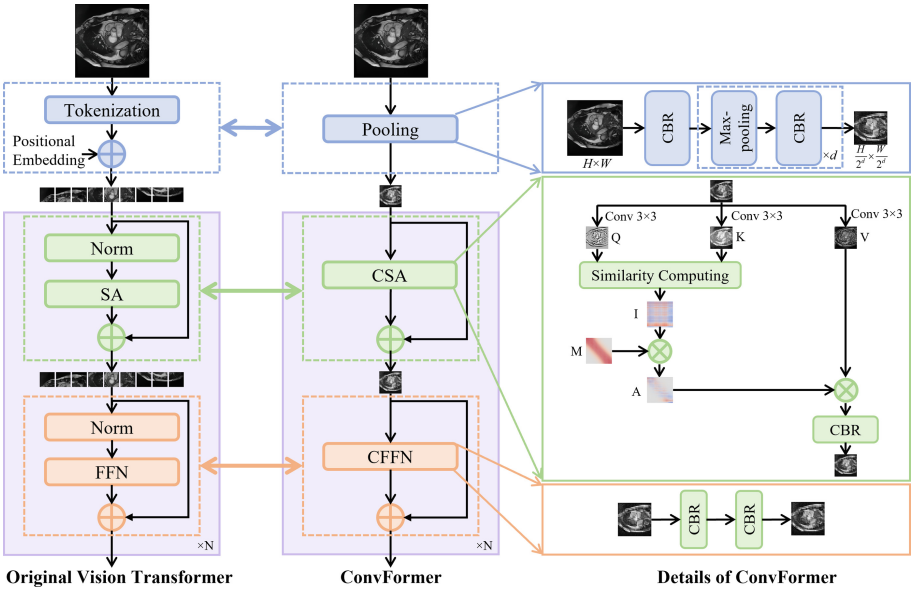

Fig. 2. Comparison between vanilla vision transformer and ConvFormer. CBR is short for the combination of convolution, batch normalization, and Relu. Multiple heads are omitted for simplicity.

3.1 Pooling vs. Tokenization

The pooling module is developed to realize the functions of tokenization (*i.e.*, making the input suitable to transformers in the channel dimension and shaping and reducing the input size when needed) while without losing details in the grid lines in tokenization. For an input $X_{in} \in \mathbb{R}^{c \times H \times W}$, convolution with a kernel size of 3×3 followed by batch normalization and Relu, is first applied to capture local features. Then, corresponding to each patch size S in ViT, total $d = \log_2 S$ downsampling operations are applied in the pooling module to produce the same resolutions. Here, each downsampling operation consists of a max-pooling with a kernel size of 2×2 and a combination of 3×3 convolution, batch normalization, and Relu. Finally, X_{in} becomes $X_1 \in \mathbb{R}^{c_m \times \frac{H}{2^d} \times \frac{W}{2^d}}$ through the pooling module where c_m is corresponding to the embedding dimension in ViT.

3.2 CNN-Style vs. Sequenced Self-attention

The building of long-range dependency in ConvFormer is relying on CNN-style self-attention, which creates an adaptive receptive field for each pixel by constructing a customized convolution kernel. Specifically, for each pixel $x_{i,j}$ of X_1, the convolution kernel $A^{i,j}$ is constructed based on two intermediate variables:

$$Q_{i,j} = \sum_{l=-1}^{1} \sum_{g=-1}^{1} E^{q}_{2+l,2+g} x_{i+l,j+g}, \tag{1}$$

$$K_{i,j} = \sum_{l=-1}^{1} \sum_{g=-1}^{1} E^{k}_{2+l,2+g} x_{i+l,j+g}, \tag{2}$$

where E^q and $E^k \in \mathbb{R}^{c_q \times c_m \times 3 \times 3}$ are the learnable projection matrices and c_q is corresponding to the embedding dimension of Q, K, and V in ViT, which incorporates the features of adjacent pixels in 3×3 neighborhood into $x_{i,j}$. Then, the initial customized convolutional kernel $I^{i,j} \in \mathbb{R}^{\frac{H}{2^d} \times \frac{W}{2^d}}$ for $x_{i,j}$ is calculated by computing the cosine similarity:

$$I^{i,j}_{m,n} = \frac{\sum_{l=0}^{c_q} Q_{i,j} K_{m,n}}{\sqrt{\sum_{l=0}^{c_q} Q^2_{i,j}} \sqrt{\sum_{l=0}^{c_q} K^2_{m,n}}}. \tag{3}$$

Here, $I^{i,j}_{m,n} \in [-1, 1]$ and seldom occurs $I^{i,j}_{m,n} = 0$. $I^{i,j}_{m,n}$ corresponds to attention score calculation in ViT (constrained to be positive while $I^{i,j}_{m,n}$ can be either positive or negative). Then, we dynamically determine the size of the customized convolution kernel for $x_{i,j}$ by introducing a learnable Gaussian distance map M:

$$M^{i,j}_{m,n} = e^{-\frac{(i-m)^2 (2^d/H)^2 + (j-n)^2 (2^d/W)^2}{2(\theta \times \alpha)^2}}, \tag{4}$$

where $\theta \in (0, 1)$ is a learnable network parameter to control the receptive field of A and α is a hyper-parameter to control the tendency of the receptive field. θ is proportional to the receptive field. For instance, under the typical setting $H = W = 256$, $d = 3$, and $\alpha = 1$, when $\theta = 0.003$, the receptive field only covers five adjacent pixels, when $\theta > 0.2$, the receptive field is global. The larger α is, the more likely A tends to have a global receptive field. Based on $I^{i,j}$ and $M^{i,j}$, $A^{i,j}$ is calculated by $A^{i,j} = I^{i,j} \times M^{i,j}$. In this way, every pixel $x^{i,j}$ has a customized size-scalable convolution kernel $A^{i,j}$. By multiplying A with V, CSA can build adaptive long-range dependency, where V can be formulated similarly according to Eq. (1). Finally, the combination of 1×1 convolution, batch normalization, and Relu is utilized to integrate features learned from long-range dependency.

3.3 Convolution Vs. Vanilla Feed-Forward Network

The convolution feed-forward network (CFFN) is to refine the features produced by CSA, just consisting of two combinations of 1×1 convolution, batch normalization, and Relu. By replacing linear projection and layer normalization in ViT, CFFN makes ConvFormer completely CNN-based, avoiding the combat between CNN and Transformer during training like CNN-Transformer hybrid approaches.

4 Experiments

4.1 Datasets and Implementation Details

ACDC[1]. A publicly-available dataset for the automated cardiac diagnosis challenge. Totally 100 scans with pixel-wise annotations of left ventricle (LV),

[1] https://www.creatis.insa-lyon.fr/Challenge/acdc/.

Table 1. Quantitative results in Dice (DSC) and Hausdorff Distance (HD).

Method	DSC (%) ↑						HD (%) ↓		
	ACDC				ISIC	ICH	ACDC	ISIC	ICH
	Avg.	RV	MYO	LV			Avg.		
SETR [5]	87.14	83.95	83.89	93.59	89.03	80.17	17.24	22.33	14.90
+Re-attention [15]	85.91	82.52	82.27	92.93	88.00	79.08	18.21	24.57	14.50
+LayerScale [21]	85.74	81.54	82.70	92.98	87.98	78.91	17.88	23.94	14.45
+Refiner [20]	85.75	83.18	81.61	92.46	86.63	78.35	18.09	25.43	14.95
+ConvFormer	**91.00**	**89.26**	**88.60**	**95.15**	**90.41***	**81.56**	**14.08**	**21.68**	**13.50**
TransUNet [12]	90.80	89.59	87.81	94.99	88.75	78.52	14.58	25.11	15.90
+Re-attention [15]	91.25	89.91	88.61	95.22	88.35	77.50	13.79	**23.15**	**15.40**
+LayerScale [21]	91.30	89.37	88.79	**95.75**	88.75	75.60	**13.68**	23.32	15.50
+Refiner [20]	90.76	88.66	88.39	95.22	87.90	76.47	14.73	25.31	15.65
+ConvFormer	**91.42**	**90.17**	**88.84**	95.25	**89.40**	**80.66**	13.96	23.19	15.70
TransFuse [13]	89.10	87.85	85.73	93.73	89.28	75.11	14.98	23.08	18.60
+Re-attention [15]	88.48	87.05	85.37	93.00	88.28	73.74	16.20	24.56	18.45
+LayerScale [21]	88.85	87.81	85.24	93.50	89.00	74.18	13.53	23.96	20.00
+Refiner [20]	89.06	87.88	85.55	93.75	85.65	74.16	14.05	26.30	18.60
+ConvFormer	**89.88**	**88.85**	**86.50**	**94.30**	**90.56***	**75.56**	**12.84**	**21.30**	**17.60**
FAT-Net [14]	91.46	90.13	88.61	95.60	89.72	83.73	13.82	22.63	16.20
+Re-attention [15]	91.61*	89.99	89.18	95.64	89.84	84.42	14.00	22.54	14.20
+LayerScale [21]	91.71*	90.01	89.39	95.71	90.06	83.87	13.50	21.93	13.70
+Refiner [20]	91.94*	90.54	**89.70**	95.58	89.20	83.14	13.37	23.35	**13.25**
+ConvFormer	**92.18***	**90.69**	89.57	**96.28**	**90.36***	**84.97**	**11.32**	**21.73**	14.10
Patcher [27]	91.41	89.56	89.12	95.53	89.11	80.54	13.55	22.16	15.70
+Re-attention [15]	91.25	89.77	88.58	95.39	89.73	79.08	14.49	21.86	17.60
+LayerScale [21]	91.07	88.94	88.81	95.46	90.16	74.13	15.48	**21.78**	18.60
+Refiner [20]	91.26	89.65	88.58	95.57	68.92	79.66	13.93	56.62	15.60
+ConvFormer	**92.07***	**90.91**	**89.54**	**95.78**	**90.18**	**81.69**	**12.29**	21.88	**15.35**

* Approaches outperforming the state-of-the-art 2D approaches on the publicly-available ACDC (*i.e.*, FAT-Net [14]: 91.46% in Avg. DSC) and ISIC (*i.e.*, Ms Red [28]: 90.25% in Avg. DSC) datasets respectively. *More comprehensive quantitative comparison results can be found in the supplemental materials.*

myocardium (MYO), and right ventricle (RV) are available [22]. Following [12,17,18], 70, 10, and 20 cases are used for training, validation, and testing respectively.

ISIC 2018[2]. A publicly-available dataset for skin lesion segmentation. Totally 2594 dermoscopic lesion images with pixel-level annotations are available [23,24]. Following [25,26], the dataset is randomly divided into 2076 images for training and 520 images for testing.

[2] https://challenge.isic-archive.com/data/.

ICH. A locally-collected dataset for hematoma segmentation. Totally 99 CT scans consisting of 2648 slices were collected and annotated by three radiologists. The dataset is randomly divided into the training, validation, and testing sets according to a ratio of 7:1:2.

Implementation Details. For a fair comparison, all the selected state-of-the-art transformer-based baselines were trained with or without ConvFormer under the same settings. All models were trained by an Adam optimizer with a learning rate of 0.0001 and a batch size of 4 for 400 rounds. Data augmentation includes random rotation, scaling, contrast augmentation, and gamma augmentation.

4.2 Results

ConvFormer can work as a plug-and-play module and replace the vanilla transformer blocks in transformer-based baselines. To evaluate the effectiveness of ConvFormer, five state-of-the-art transformer-based approaches are selected as backbones, including SETR [5], TransUNet [12], TransFuse [13], FAT-Net [14], and Patcher [27]. SETR and Patcher utilize pure-transformer encoders, while TransUNet, TransFuse, and FAT-Net adopt CNN-Transformer hybrid encoders. In addition, three state-of-the-art methods for addressing attention collapse, including Re-attention [15], LayerScale [21], and Refiner [20], are equipped with the above transformer-based baselines for comparison.

Quantitative Results. Quantitative results of ConvFormer embedded into various transformer-based baselines on the three datasets are summarized in Table 1. ConvFormer achieves consistent performance improvements on all five backbones. Compared to CNN-Transformer hybrid approaches (*i.e.*, TransUNet, TransFuse, and FAT-Net), ConvFormer is more beneficial on pure-transformer approaches (*i.e.*, SETR and Patcher). Specifically, with ConvFormer, SETR achieves an average increase of 3.86%, 1.38%, and 1.39% in Dice on the ACDC, ISIC, and ICH datasets respectively, while the corresponding performance improvements of Patcher are 0.66%, 1.07%, and 1.15% respectively. Comparatively, in CNN-Transformer hybrid approaches, as analyzed above, CNNs would be more dominating against transformers during training. Despite this, re-balancing CNNs and Transformers through ConvFormer can build better long-range dependency for consistent performance improvement.

Comparison with SOTA Approaches. Quantitative results compared with the state-of-the-art approaches to addressing attention collapse are summarized in Table 1. In general, given relatively limited training data, existing approaches designed for natural image processing are unsuitable for medical image segmentation, resulting in unstable performance across different backbones and datasets. Comparatively, ConvFormer consistently outperforms these approaches and brings stable performance improvements to various backbones across datasets, demonstrating the excellent generalizability of ConvFormer as a plug-and-play module.

Visualization of Self-Attention Matrices. To qualitatively evaluate the effectiveness of ConvFormer in addressing attention collapse and building efficient long-range dependency, we visualize the self-attention matrices with and

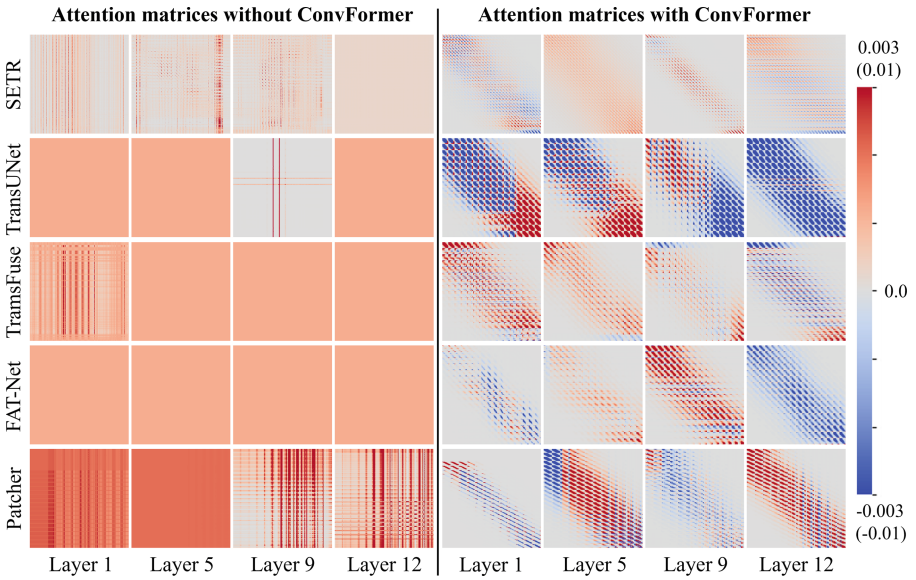

Fig. 3. Visualization of self-attention matrices by baselines w/ and w/o ConvFormer.

Table 2. Ablation study of hyper-parameter α on the ACDC dataset.

α	0.2	0.4	0.6	0.8	1.0
Dice (%)	90.71	**91.00**	90.76	90.66	90.45

without ConvFormer as illustrated in Fig. 3. By introducing ConvFormer, attention collapse is effectively alleviated. Compare to the self-attention matrices of baselines, the matrices learned by ConvFormer are more diverse. Specifically, the interactive range for each pixel is scalable, being small for locality preserving or being large for global receptive fields. Besides, dependency is no longer constrained to be positive like ViT, which is more consistent with convolution kernels. *Qualitative segmentation results of different approaches on the three datasets can be found in the supplemental materials.*

Ablation Study As described in Sec. 3.2, α is to control the receptive field tendency in ConvFormer, The larger the α, the more likely ConvFormer contains larger receptive fields. To validate this, we conduct an ablation study on α as summarized in Table 2. In general, using a large α does not necessarily lead to more performance improvements, which is consistent with our observation that not every pixel needs global information for segmentation.

5 Conclusions

In this paper, we construct the transformer as a kernel-scalable convolution to address the attention collapse and build diverse long-range dependencies for

efficient medical image segmentation. Specifically, it consists of pooling, CNN-style self-attention (CSA), and convolution feed-forward network (CFFN). The pooling module is first applied to extract the locality details while reducing the computational costs of the following CSA module by downsampling the inputs. Then, CSA is developed to build adaptive long-range dependency by constructing CSA as a kernel-scalable convolution, Finally, CFFN is used to refine the features of each pixel. Experimental results on five state-of-the-art baselines across three datasets demonstrate the prominent performance of ConvFormer, stably exceeding the baselines and comparison methods across three datasets.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant 62271220 and Grant 62202179, and in part by the Natural Science Foundation of Hubei Province of China under Grant 2022CFB585. The computation is supported by the HPC Platform of HUST.

References

1. Vaswani, A., et al.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
2. 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
3. Schlemper, J., et al.: Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. **53**, 197–207 (2019)
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6881–6890 (2021)
6. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16000–16009 (2022)
7. You, C., et al.: Class-aware generative adversarial transformers for medical image segmentation. arXiv preprint arXiv:2201.10737 (2022)
8. Karimi, D., Vasylechko, S.D., Gholipour, A.: Convolution-free medical image segmentation using transformers. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 78–88. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_8
9. Zhang, Y., et al.: mmFormer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: Wang, Li., Dou, Q., Fletcher, P.T., Speidel S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13431, pp. 107–117. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_11
10. Wang, Z., et al.: SMESwin unet: merging CNN and transformer for medical image segmentation. In: Wang, Li., Dou, Q., Fletcher, P.T., Speidel S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13431, pp. 517–526. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_50

11. Li, H., Chen, L., Han, H., Zhou, S.K.: SATr: slice attention with transformer for universal lesion detection. In: Wang, Li., Dou, Q., Fletcher, P.T., Speidel S., Li, S. (eds.) MICCAI 2022, LNCS, vol. 13431, pp. 163–174. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_16
12. Chen, J., et al. Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
13. Zhang, Y., Liu, H., Hu, Q.: Transfuse: fusing transformers and CNNs for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 14–24. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_2
14. 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)
15. Zhou, D., et al.: DeepViT: towards deeper vision transformer. arXiv preprint arXiv:2103.11886 (2021)
16. Huang, X., Deng, Z., Li, D., Yuan, X.: Missformer: an effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162 (2021)
17. Cao, H., et al. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
18. Xu, G., Wu, X., Zhang, X., He, X.: Levit-unet: make faster encoders with transformer for medical image segmentation. arXiv preprint arXiv:2107.08623 (2021)
19. Liu, W., et al.: Phtrans: parallelly aggregating global and local representations for medical image segmentation. In: Wang, Li., Dou, Q., Fletcher, P.T., Speidel S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13431, pp. 235–244. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_23
20. Zhou, D., et al.: Refiner: refining self-attention for vision transformers. arXiv preprint arXiv:2106.03714 (2021)
21. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 32–42 (2021)
22. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imag. **37**(11), 2514–2525 (2018)
23. 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)
24. 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)
25. Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., Zhang, D.: Ds-transunet: dual swin transformer u-net for medical image segmentation. IEEE Instrum. Meas. **71**, 1–15 (2022)
26. Chen, B., Liu, Y., Zhang, Z., Lu, G., Kong, A.W.K.: Transattunet: multi-level attention-guided u-net with transformer for medical image segmentation. arXiv preprint arXiv:2107.05274 (2021)
27. Ou, Y., et al.: Patcher: patch transformers with mixture of experts for precise medical image segmentation. In: Wang, Li., Dou, Q., Fletcher, P.T., Speidel S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13431, pp. 475–484. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_46
28. Dai D., et al.: Ms RED: a novel multi-scale residual encoding and decoding network for skin lesion segmentation. Med. Image Anal. **75**, 102293 (2022)