Laplacian-Former: Overcoming the Limitations of Vision Transformers in Local Texture Detection

Reza Azad[1(✉)], Amirhossein Kazerouni[2], Babak Azad[3], Ehsan Khodapanah Aghdam[4], Yury Velichko[5], Ulas Bagci[5], and Dorit Merhof[6]

[1] Faculty of Electrical Engineering and Information Technology, RWTH Aachen University, Aachen, Germany
azad@pc.rwth-aachen.de
[2] School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran
[3] South Dakota State University, Brookings, USA
[4] Department of Electrical Engineering, Shahid Beheshti University, Tajrish, Iran
[5] Department of Radiology, Northwestern University, Chicago, USA
[6] Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany

Abstract. Vision Transformer (ViT) models have demonstrated a breakthrough in a wide range of computer vision tasks. However, compared to the Convolutional Neural Network (CNN) models, it has been observed that the ViT models struggle to capture high-frequency components of images, which can limit their ability to detect local textures and edge information. As abnormalities in human tissue, such as tumors and lesions, may greatly vary in structure, texture, and shape, high-frequency information such as texture is crucial for effective semantic segmentation tasks. To address this limitation in ViT models, we propose a new technique, Laplacian-Former, that enhances the self-attention map by adaptively re-calibrating the frequency information in a Laplacian pyramid. More specifically, our proposed method utilizes a dual attention mechanism via efficient attention and frequency attention while the efficient attention mechanism reduces the complexity of self-attention to linear while producing the same output, selectively intensifying the contribution of shape and texture features. Furthermore, we introduce a novel efficient enhancement multi-scale bridge that effectively transfers spatial information from the encoder to the decoder while preserving the fundamental features. We demonstrate the efficacy of Laplacian-former on multi-organ and skin lesion segmentation tasks with +1.87% and +0.76% dice scores compared to SOTA approaches, respectively. Our implementation is publically available at GitHub.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_70.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 736–746, 2023.
https://doi.org/10.1007/978-3-031-43898-1_70

Keywords: Deep Learning · Texture · Segmentation · Laplacian Transformer

1 Introduction

The recent advancements in Transformer-based models have revolutionized the field of natural language processing and have also shown great promise in a wide range of computer vision tasks [5]. As a notable example, the Vision Transformer (ViT) model utilizes Multi-head Self-Attention (MSA) blocks to globally model the interactions between semantic tokens created by treating local image patches as individual elements [7]. This approach stands in contrast to CNNs, which hierarchically increase their receptive field from local to global to capture a global semantic representation. Nevertheless, recent studies [3,20] have shown that ViT models struggle to capture high-frequency components of images, which can limit their ability to detect local textures and it is vital for many diagnostic and prognostic tasks. This weakness in local representation can be attributed to the way in which ViT models process images. ViT models split an image into a sequence of patches and model their dependencies using a self-attention mechanism, which may not be as effective as the convolution operation used in CNN models in extracting local features within receptive fields. This difference in how ViT and CNN process images may explain the superior performance of CNN models in local feature extraction [1,8]. Innovative approaches have been proposed in recent years to address the insufficient local texture representation within Transformer models. One such approach is the integration of CNN and ViT features through complementary methods, aimed at seamlessly blending the strengths of both in order to compensate for any shortcomings in local representation [5].

Transformers as a Complement to CNNs: TransUNet [5] is one of the earliest approaches incorporating the Transformer layers into the CNN bottleneck to model both local and global dependency using the combination of CNN and ViT models. Heidari et al. [11] proposed a novel solution called HiFormer, which leverages a Swin Transformer module and a CNN-based encoder to generate two multi-scale feature representations, which are then integrated via a Double-Level Fusion module. UNETR [10] used a Transformer to create a powerful encoder with a CNN decoder for 3D medical image segmentation. By bridging the CNN-based encoder and decoder with the Transformer, CoTr [26], and TransBTS [22], the segmentation performance in low-resolution stages was improved. Despite these advances, there remain some limitations in these methods such as computationally inefficiency (e.g., TransUNet model), the requirement of a heavy CNN backbone (e.g., HiFormer), and the lack of consideration for multi-scale information. These limitations have resulted in less effective network learning results in the field of medical image segmentation.

New Attention Models: The redesign of the self-attention mechanism within pure Transformer models is another method aiming to augment feature repre-

sentation to enhance the local feature representation ultimately. In this direction, Swin-Unet [4] utilizes a linear computational complexity Swin Transformer [14] block in a U-shaped structure as a multi-scale backbone. MISSFormer [12] besides exploring the Efficient Transformer [25] counterpart to diminish the parameter overflow of vision transformers, applies a non-invertible downsampling operation on input blocks transformer to reduce the parameters. D-Former [24] is a pure transformer-based pipeline that comprises a double attention module to capture locally fine-grained attention and interaction with different units in a dilated manner through its mechanism.

Drawbacks of Transformers: Recent research has revealed that traditional self-attention mechanisms, while effective in addressing local feature discrepancies, have a tendency to overlook important high-frequency information such as texture and edge details [21]. This is especially problematic for tasks like tumor detection, cancer-type identification through radiomics analysis, as well as treatment response assessment, where abnormalities often manifest in texture. Moreover, self-attention mechanisms have a quadratic computational complexity and may produce redundant features [18].

Our Contributions: ❶ We propose Laplacian-Former, a novel approach that includes new efficient attention (EF-ATT) consisting of two sub-attention mechanisms: *efficient attention* and *frequency attention*. The efficient attention mechanism reduces the complexity of self-attention to linear while producing the same output. The frequency attention mechanism is modeled using a Laplacian pyramid to emphasize each frequency information's contribution selectively. Then, a parametric frequency attention fusion strategy to balance the importance of shape and texture features by recalibrating the frequency features. These two attention mechanisms work in parallel. ❷ We also introduce a novel efficient enhancement multi-scale bridge that effectively transfers spatial information from the encoder to the decoder while preserving the fundamental features. ❸ Our method not only alleviates the problem of the traditional self-attention mechanism mentioned above, but also it surpasses all its counterparts in terms of different evaluation metrics for the tasks of medical image segmentation.

2 Methods

In our proposed network, illustrated in Fig. 1, taking an input image $X \in R^{H \times W \times C}$ with spatial dimensions H and W, and C channels, it is first passed through a patch embedding module to obtain overlapping patch tokens of size 4×4 from the input image. The proposed model comprises four encoder blocks, each containing two efficient enhancement Transformer layers and a patch merging layer that downsamples the features by merging 2×2 patch tokens and increasing the channel dimension. The decoder is composed of three efficient enhancement Transformer blocks and four patch-expanding blocks, followed by a segmentation head to retrieve the final segmentation map. Laplacian-Former then employs a novel efficient enhancement multi-scale bridge to capture local

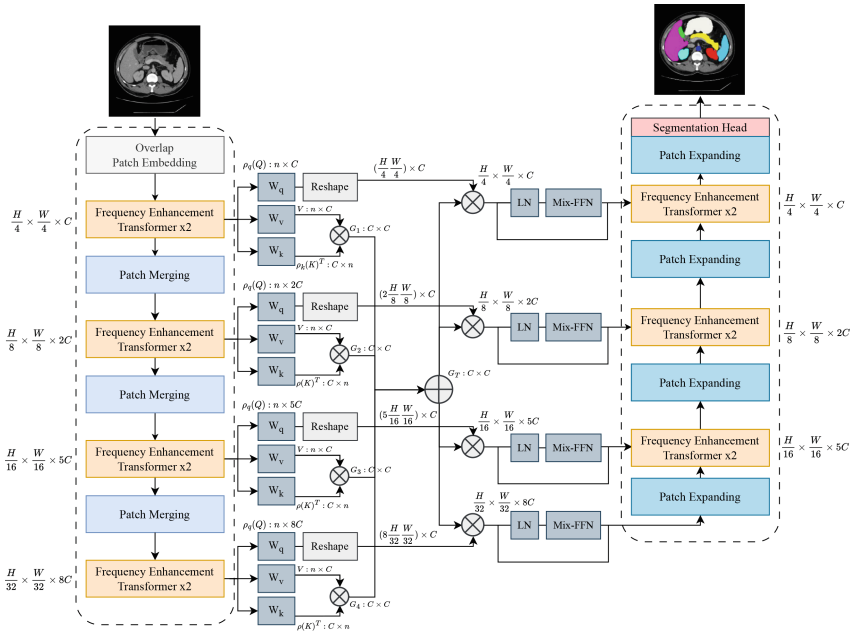


Fig. 1. Architecture of our proposed Laplacian-Former.

and global correlations of different scale features and effectively transfer the underlying features from the encoder to the decoder.

2.1 Efficient Enhancement Transformer Block

In medical imaging, it is important to distinguish different structures and tissues, especially when tissue boundaries are ill-defined. This is often the case for accurate segmentation of small abnormalities, where high-frequency information plays a critical role in defining boundaries by capturing both textures and edges. Inspired by this, we propose an Efficient Enhancement Transformer Block that incorporates an Efficient Frequency Attention (EF-ATT) mechanism to capture contextual information of an image while recalibrating the representation space within an attention mechanism and recovering high-frequency details.

Our efficient enhancement Transformer block first takes a LayerNorm (LN) from the input x. Then it applies the EF-ATT mechanism to capture contextual information and selectively include various types of frequency information while using the Laplacian pyramid to balance the importance of shape and texture features. Next, x and diversity-enhanced shortcuts are added to the output of the attention mechanism to increase the diversity of features. It is proved in [19] that as Transformers become deeper, their features become less varied, which restrains their representation capacity and prevents them from attaining optimal performance. To address this issue, we have implemented an *augmented short-*

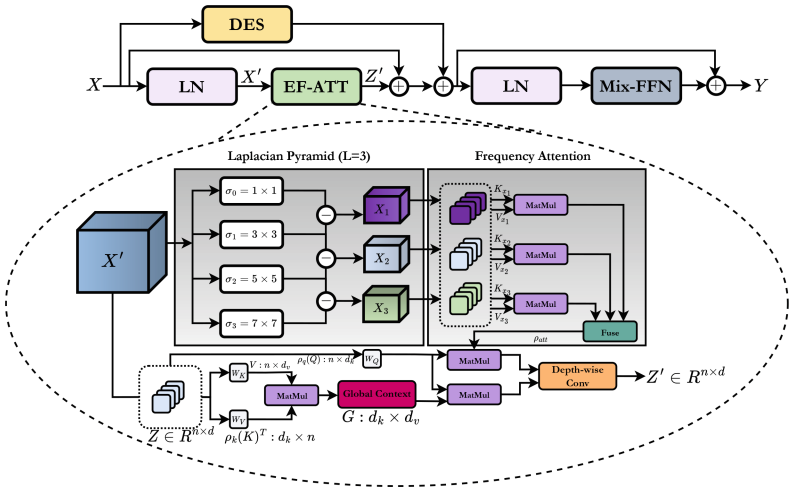


Fig. 2. The structure of our frequency enhancement Transformer block.

cut method from [9], a Diversity-Enhanced Shortcut (DES), employing a Kronecker decomposition-based projection. This approach involves inserting additional paths with trainable parameters alongside the original shortcut x, which enhances feature diversity and improves performance while requiring minimal hardware resources. Finally, we apply LayerNorm and MiX-FFN [25] to the resulting feature representation to enhance its power. This final step completes our efficient enhancement Transformer block, as illustrated in Fig. 2.

2.2 Efficient Frequency Attention (EF-ATT)

The traditional self-attention block computes the attention score S using query ($\mathbf{Q}$) and key ($\mathbf{K}$) values, normalizes the result using Softmax, and then multiplies the normalized attention map with value ($\mathbf{V}$):

$$S(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\left(\frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{1}$$

where d_k is the embedding dimension. One of the main limitations of the dot-product mechanism is that it generates redundant information, resulting in unnecessary computational complexity. Shen et al. [18] proposed to represent the context more effectively by reducing the computational burden from $\mathcal{O}(n^2)$ to linear form $\mathcal{O}(d^2 n)$:

$$E(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \rho_{\mathbf{q}}(\mathbf{Q})\left(\rho_{\mathbf{k}}(\mathbf{K})^{\mathbf{T}}\mathbf{V}\right). \tag{2}$$

Their approach involves applying the Softmax function (ρ) to the key and query vectors to obtain normalized scores and formulating the global context by

multiplying the key and value matrix. They demonstrate that efficient attention E can provide an equivalent representation of self-attention while being computationally efficient. By adopting this approach, we can alleviate the issues of feature redundancy and computational complexity associated with self-attention.

Wang et al. [21] explored another major limitation of the self-attention mechanism, where they demonstrated through theoretical analysis that self-attention operates as a low-pass filter that erases high-frequency information, leading to a loss of feature expressiveness in the model's deep layers. Authors found that the Softmax operation causes self-attention to keep low-frequency information and loses its fine details. Motivated by this, we propose a new frequency recalibration technique to address the limitations of self-attention, which only focuses on low-frequency information (which contains shape information) while ignoring the higher frequencies that carry texture and edge information. First, we construct a Laplacian pyramid to determine the different frequency levels of the feature maps. The process begins by extracting $(L+1)$ Gaussian representations from the encoded feature using different variance values of the Gaussian function:

$$\mathbf{G}_l(\mathbf{X}) = \mathbf{X} * \frac{1}{\sigma_l\sqrt{2\pi}} e^{-\frac{i^2+j^2}{2\sigma_l^2}}, \tag{3}$$

where $\mathbf{X}$ refers to the input feature map, (i, j) corresponds to the spatial location within the encoded feature map, the variable σ_l denotes the variance of the Gaussian function for the l-th scale, and the symbol $*$ represents the convolution operator. The pyramid is then built by subtracting the l-th Gaussian function $(\mathbf{G}_l)$ output from the $(l+1)$-th output $(\mathbf{G}_l - \mathbf{G}_{l+1})$ to encode frequency information at different scales. The Laplacian pyramid is composed of multiple levels, each level containing distinct types of information. To ensure a balanced distribution of low and high-frequency information in the model, it is necessary to efficiently aggregate the features from all levels of the frequency domain. Hence, we present frequency attention that involves multiplying the key and value of each level $(\mathbf{X}_l)$ to calculate the attention score and then fuses the resulting attention scores of all levels using a fusion module, which performs summation. The resulting attention score is multiplied by Query $(\mathbf{Q})$ to obtain the final frequency attention result, which subsequently concatenates with the efficient attention result and applies the depth-wise convolution with the kernel size of $2 \times 1 \times 1$ in order to aggregate both information and recalibrate the feature map, thus allowing for the retrieval of high-frequency information.

2.3 Efficient Enhancement Multi-scale Bridge

It is widely known that effectively integrating multi-scale information can lead to improved performance [12]. Thus, we introduce the Efficient Enhancement Multi-scale Bridge as an alternative to simply concatenating the features from the encoder and decoder layers. The proposed bridge, depicted in Fig. 1, delivers spatial information to each decoder layer, enabling the recovery of intricate details while generating output segmentation masks. In this approach, we aim

to calculate the efficient attention mechanism for each level and fuse the multi-scale information in their context; thus, it is important that all levels' embedding dimension is of the same size. Therefore, in order to calculate the global context ($\mathbf{G}_i$), we parametrize the query and value of each level using a convolution 1×1 where it gets the size of mC and outputs C, where m equals 1, 2, 5, and 8 for the first to fourth levels, respectively. We multiply the new key and value to each other to attain the global context. We then use a summation module to aggregate the global context of all levels and reshape the query for matrix multiplication with the augmented global context. Taking the second level with the dimension of $\frac{H}{8} \times \frac{W}{8} \times 2C$, the key and value are mapped to $(\frac{H}{8}\frac{W}{8}) \times C$, and the query to $(2\frac{H}{8}\frac{W}{8}) \times C$. The augmented global context with the shape of $C \times C$ is then multiplied by the query, resulting in an enriched feature map with the shape of $(2\frac{H}{8}\frac{W}{8}) \times C$. We reshape the obtained feature map into $\frac{H}{8} \times \frac{W}{8} \times 2C$ and feed it through an LN and MiX-FFN module with a skip connection to empower the feature representations. The resulting output is combined with the expanded feature map, and then projected using a linear layer onto the same size as the encoder block corresponding to that level.

3 Results

Our proposed technique was developed using the PyTorch library and executed on a single RTX 3090 GPU. A batch size of 24 and a stochastic gradient descent algorithm with a base learning rate of 0.05, a momentum of 0.9, and a weight decay of 0.0001 was utilized during the training process, which was carried out for 400 epochs. For the loss function, we used both cross-entropy and Dice losses ($Loss = \gamma \cdot L_{dice} + (1 - \gamma) \cdot L_{ce}$), γ set to 0.6 empirically.

Datasets: We tested our model using the *Synapse* dataset [13], which comprises 30 cases of contrast-enhanced abdominal clinical CT scans (a total of 3,779 axial slices). Each CT scan consists of $85 \sim 198$ slices of the in-plane size of 512×512 and has annotations for eight different organs. We followed the same preferences for data preparation analogous to [5]. We also followed [2] experiments to evaluate our method on the ISIC 2018 skin lesion dataset [6] with 2,694 images.

Table 1. Comparison results of the proposed method on the *Synapse* dataset. Blue indicates the best result, and red indicates the second-best.

Methods	# Params (M)	DSC ↑	HD ↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
R50 U-Net [5]	30.42	74.68	36.87	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
U-Net [16]	14.8	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Att-UNet [17]	34.9	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransUNet [5]	105.28	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet [4]	27.17	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
LeVit-Unet [27]	52.17	78.53	16.84	78.53	62.23	84.61	80.25	93.11	59.07	88.86	72.76
TransDeepLab [2]	21.14	80.16	21.25	86.04	69.16	84.08	79.88	93.53	61.19	89.00	78.40
HiFormer [11]	25.51	80.39	14.70	86.21	65.69	85.23	79.77	94.61	59.52	90.99	81.08
EffFormer	22.31	80.79	17.00	85.81	66.89	84.10	81.81	94.80	62.25	91.05	79.58
LaplacianFormer (without bridge)	23.87	81.59	17.31	87.41	69.57	85.22	80.46	94.68	63.71	91.47	78.23
LaplacianFormer	27.54	81.90	18.66	86.55	71.19	84.23	80.52	94.90	64.75	91.91	81.14

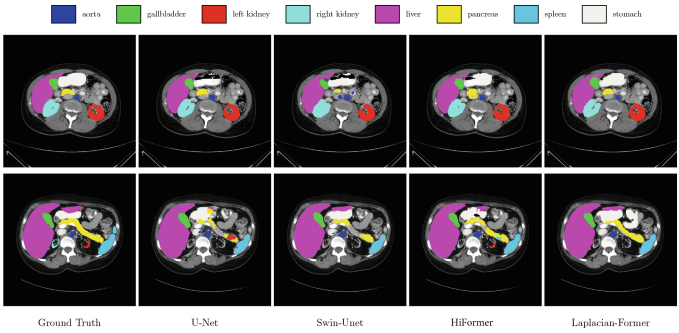


Fig. 3. Segmentation results of the proposed method on the *Synapse* dataset. Our Laplacian-Former shows finer boundaries (high-frequency details) for the region of the stomach and less false positive prediction for the pancreas.

Synapse Multi-organ Segmentation: Table 1 presents a comparison of our proposal with previous SOTA methods using the DSC and HD metrics across eight abdominal organs. Laplacian-Former clearly outperforms SOTA CNN-based methods. We extensively evaluated EfficientFormer (EffFormer) plus another drift of Laplacian-Former without utilizing the bridge connections to endorse the superiority of Laplacian-Former. Laplacian-Former exhibits superior learning ability on the Dice score metric compared to other transformer-based models, achieving an increase of +1.59% and +2.77% in Dice scores compared to HiFormer and Swin-Unet, respectively. Figure 3 illustrates a qualitative result of our method for different organ segmentation, specifically we can observe that the LalacianFormer produces a precise boundary segmentation on Gallbladder, Liver, and Stomach organs. It is noteworthy to mention that our pipeline, as a pure transformer-based architecture trained from scratch without pretraining weights, outperforms all previously presented network architectures.

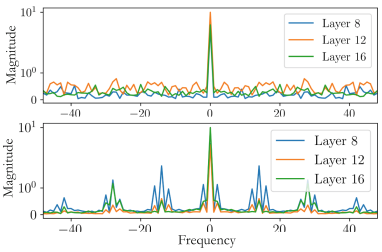
Skin Lesion Segmentation: Table 2a shows the comparison results of our proposed method, Laplacian-Former, against leading methods on the skin lesion segmentation benchmark. Our approach outperforms other competitors across most evaluation metrics, indicating its excellent generalization ability across different datasets. In particular, our approach performs better than hybrid methods such as TMU-Net [15] and pure transformer-based methods such as Swin-Unet [4]. Our method achieves superior performance by utilizing the frequency attention in a pyramid scale to model local textures. Specifically, our frequency attention emphasizes the fine details and texture characteristics that are indicative of skin lesion structures and amplifies regions with significant intensity variations, thus accentuating the texture patterns present in the image and resulting in better performance. In addition, we provided the spectral response of LaplacianFormer vs. Standard Transformer in identical layers in Table 2b. It is evident Standard design frequency response in deep layers of structure attenuates more than the LaplacianFormer, which is a visual endorsement of the capability of Laplacian-

Table 2. (a) Performance comparison of Laplacian-Former against the SOTA approaches on *ISIC 2018* skin lesion datset. Blue and red indicates the best and the second-best results. (b) Frequency response analysis on the LaplacianFormer (up) vs. Standard Transformer (down).

(a) *ISIC 2018* dataset

Methods	ISIC 2018			
	DSC	SE	SP	ACC
U-Net [17]	0.8545	0.8800	0.9697	0.9404
Att-UNet [18]	0.8566	0.8674	0.9863	0.9376
TransUNet [5]	0.8499	0.8578	0.9653	0.9452
FAT-Net [24]	0.8903	0.9100	0.9699	0.9578
TMU-Net [15]	0.9059	0.9038	0.9746	0.9603
Swin-Unet [4]	0.8946	0.9056	0.9798	0.9605
EffFormer	0.8909	0.9034	0.9701	0.9579
Laplacian-Former (without bridge)	0.9100	0.9289	0.9655	0.9611
Laplacian-Former	0.9128	0.9290	0.9715	0.9626

(b) Spectral Response



Former for its ability to preserve high-frequency details. The supplementary provides more visualization results.

4 Conclusion

In this paper, we introduce Laplacian-Former, a novel standalone transformer-based U-shaped architecture for medical image analysis. Specifically, we address the transformer's inability to capture local context as high-frequency details, e.g., edges and boundaries, by developing a new design within a scaled dot attention block. Our pipeline benefits the multi-resolution Laplacian module to compensate for the lack of frequency attention in transformers. Moreover, while our design takes advantage of the efficiency of transformer architectures, it keeps the parameter numbers low.

References

1. Azad, R., Fayjie, A.R., Kauffmann, C., Ben Ayed, I., Pedersoli, M., Dolz, J.: On the texture bias for few-shot CNN segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2674–2683 (2021)
2. Azad, R., et al.: Transdeeplab: convolution-free transformer-based DeepLab v3+ for medical image segmentation. In: Rekik, I., Adeli, E., Park, S.H., Cintas, C. (eds.) PRIME 2022. LNCS, vol. 13564, pp. 91–102. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16919-9_9
3. Bai, J., Yuan, L., Xia, S.T., Yan, S., Li, Z., Liu, W.: Improving vision transformers by revisiting high-frequency components. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13684, pp. 1–18. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20053-3_1
4. Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13803, pp. 205–218. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_9

5. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
6. 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)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=YicbFdNTTy
8. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (2018)
9. Gu, J., et al.: Multi-scale high-resolution vision transformer for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12094–12103 (2022)
10. Hatamizadeh, A., et al.: Unetr: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
11. 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)
12. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: an effective transformer for 2D medical image segmentation. IEEE Trans. Med. Imaging (2022). https://doi.org/10.1109/TMI.2022.3230943
13. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: MICCAI multi-atlas labeling beyond the cranial vault-workshop and challenge. In: Proceedings of MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge, vol. 5, p. 12 (2015)
14. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
15. Reza, A., Moein, H., Yuli, W., Dorit, M.: Contextual attention network: transformer meets U-net. arXiv preprint arXiv:2203.01932 (2022)
16. 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
17. Schlemper, J., et al.: Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. **53**, 197–207 (2019)
18. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: attention with linear complexities. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3531–3539 (2021)
19. Tang, Y., et al.: Augmented shortcuts for vision transformers. Adv. Neural. Inf. Process. Syst. **34**, 15316–15327 (2021)
20. Wang, P., Zheng, W., Chen, T., Wang, Z.: Anti-oversmoothing in deep vision transformers via the fourier domain analysis: from theory to practice. In: International Conference on Learning Representations (2022)
21. Wang, P., Zheng, W., Chen, T., Wang, Z.: Anti-oversmoothing in deep vision transformers via the fourier domain analysis: from theory to practice. In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=O476oWmiNNp

22. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: multimodal brain tumor segmentation using transformer. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 109–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_11
23. 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)
24. Wu, Y., et al.: D-former: a U-shaped dilated transformer for 3D medical image segmentation. Neural Comput. Appl. 1–14 (2022)
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. Xie, Y., Zhang, J., Shen, C., Xia, Y.: CoTr: efficiently bridging CNN and transformer for 3D medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 171–180. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_16
27. 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)