Memory transformers for full context and high-resolution 3D Medical Segmentation

Loic Themyr[1,2][0000−0003−1396−2383], Clément Rambour[1][0000−0002−9899−3201], Nicolas Thome[1][0000−0003−4871−3045], Toby Collins[2][0000−0002−9441−8306], and Alexandre Hostettler[2][0000−0001−8269−6766]

[1] Conservatoire National des Arts et Métiers, Paris 75014, France
[2] IRCAD, Strasbourg 67000, France
loic.themyr@lecnam.net

Abstract. Transformer models achieve state-of-the-art results for image segmentation. However, achieving long-range attention, necessary to capture global context, with high-resolution 3D images is a fundamental challenge. This paper introduces the Full resolutIoN mEmory (FINE) transformer to overcome this issue. The core idea behind FINE is to learn memory tokens to indirectly model full range interactions while scaling well in both memory and computational costs. FINE introduces memory tokens at two levels: the first one allows full interaction between voxels within local image regions (patches), the second one allows full interactions between all regions of the 3D volume. Combined, they allow full attention over high resolution images, *e.g.* 512 x 512 x 256 voxels and above. Experiments on the BCV image segmentation dataset shows better performances than state-of-the-art CNN and transformer baselines, highlighting the superiority of our full attention mechanism compared to recent transformer baselines, *e.g.* CoTr, and nnFormer.

Keywords: Transformers · 3D segmentation · full context, high-resolution.

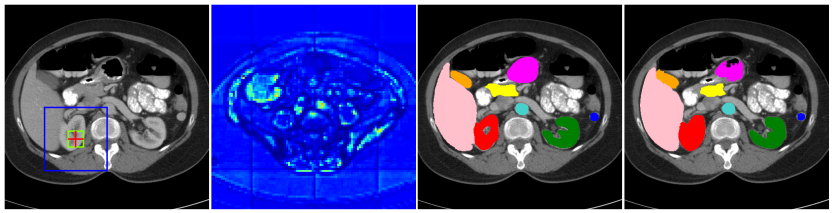
1 Introduction

Convolutional encoder-decoder models have achieved remarkable performance for medical image segmentation [1, 10]. U-Net [24] and other U-shaped architectures remain popular and competitive baselines. However, the receptive fields of these CNNs are small, both in theory and in practice [17], preventing them from exploiting global context information.

Transformers witnessed huge successes for natural language processing [26, 4] and recently in vision for image classification [5]. One key challenge in 3D semantic segmentation is their scalability, since attention's complexity is quadratic with respect to the number of inputs.

Efficient attention mechanisms have been proposed, including sparse or low-rank attention matrices [21, 28], kernel-based methods [20, 12], window [16, 6], and memory transformers [22, 14]. Multi-resolution transformers [16, 29, 30] apply attention in a hierarchical manner by chaining multiple window transform-

ers. Attention at the highest resolution level is thus limited to local image sub-windows. The receptive field is gradually increased through pooling operations. Multi-resolution transformers have recently shown impressive performances for various 2D medical image segmentation tasks such as multi-organ [11, 25, 2], histopathological [15], skin [27], or brain [23] segmentation.



a) Input image b) FINE attention c) Ground truth d) FINE prediction

Fig. 1. Proposed full resolution memory transformer (FINE). To segment the kidney voxel in a) (red cross), FINE combines high-resolution and full contextual information, as shown in the attention map in b). This is in contrast to nnFormer [33] (resp. CoTr [31]), which receptive field is limited to the green (resp. blue) region in a). FINE thus properly segments the organs, as show in d).

Recent attempts have been made to apply transformers for 3D medical image segmentation. nnFormer [33] is a 3D extension of SWIN [16] with a U-shape architecture. One limitation relates to the inherent compromise in multi-resolution, which prevents it from jointly using global context and high-resolution information. In [33], only local context is leveraged in the highest-resolution features maps. Models using deformable transformers such as CoTr [31] are able to leverage sparse global-context. A strong limitation shared by nnFormer and CoTr is that they cannot process large volumes at once and must rely on training the segmentation model on local 3D random crops. Consequently, full global contextual information is unavailable and positional encoding can be meaningless. On BCV [13], cropped patch size is about $128 \times 128 \times 64$ which only covers about 6% of the original volume.

This paper introduces the Full resolutIoN mEmory (FINE) transformer. This is, to the best of our knowledge, the first attempt at processing full-range interactions at all resolution levels with transformers for 3D medical image segmentation. To achieve this goal, memory tokens are used to indirectly enable full-range interactions between all volume elements, even when training with 3D crops. Inside each 3D crop, FINE introduces memory tokens associated to local windows. A second level of localized memory is introduced at the volume level to enable full interactions between all 3D volume patches. We show that FINE outperforms state-of-the-art CNN, transformers, and hybrid methods on the 3D multi-organ BCV dataset [13]. Fig. 1 illustrates the rationale of FINE to segment the red crossed kidney voxel in a). We can see that FINE's attention map covers the whole image, enabling to model long-range interactions between

organs. In contrast, the receptive field of state-of-the art methods only cover a small portion of the volume, *e.g.* the crop size (blue) for CoTr [31] or the even smaller window's size (green) for nnFormer [33] at the highest resolution level.

2 FINE transfomer

In this section, we detail the FINE transformer for 3D segmentation of medical images leveraging global context and full resolution information, as shown in Fig. 2. FINE is generic and can be added to most multi-resolution transformer backbones [16, 2, 33]. We chose to incorporate it to nnFormer [33], a strong model for 3D segmentation (see supplementary material).

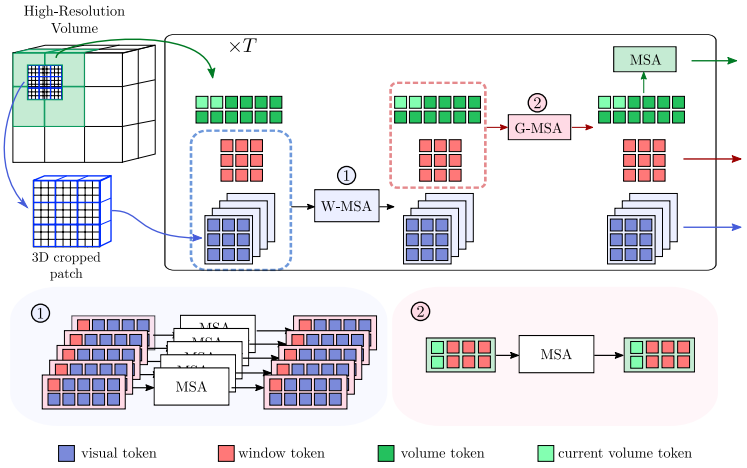


Fig. 2. To segment the cropped patch in blue and model global context, two level of memory tokens are introduced: window (red) and volume (green) tokens. First, the blue crop is divided into windows over which Multi-head Self-Attention (MSA) is performed in parallel. For each window, the sequence of visual tokens (blue) is augmented with a specific window token. Second, the local information embedded into each window token is shared between all window tokens and volume tokens intersecting with the crop (light green). Finally, high-level information is shared between all volume tokens).

2.1 Memory tokens for high resolution semantic segmentation

The core idea in FINE is to introduce memory tokens to enable full-range interactions between all voxels at all resolution levels with random cropping. We introduce memory tokens at two levels.

Window tokens. Multiple memory tokens can represent embeddings specific to regions of the feature maps [8]. Sharing these representations can thus leverage the small receptive field associated with window transformers' early

stages. In this optic, we add specific memory tokens to the sequence of visual tokens associated with each window. We chose to call them window tokens to avoid any confusion.

Volume tokens. When dealing with high-resolution volumes, random cropping is a common training strategy. This approach is a source of limitation as only a portion of the spatial context is known by the model. Worse, no efficient positional embedding can be injected as the model has no complete knowledge of the body structure. Our memory tokens overcome this issue by keeping track of the observed part of the volume. These volume tokens are associated with each element of a grid covering the entire volume and called by the transformer blocs when performing the segmentation of a cropped patch. As can be seen Fig. 4, the volume tokens induce a positional encoding learned over the entire volume.

Discussion on memory tokens. The window and volume tokens can be seen as a generalisation of the class tokens used in NLP or image classification [4, 5, 32]. In image classification, one class token is used as a global representation of the input and sent to the classifier. In semantic segmentation, more local information needs to be preserved which requires more memory tokens.

2.2 Memory based global context

Each level of memory token in FINE is related to a subdivision of the input. These memory tokens and their corresponding regions are illustrated in Fig. 2. The high-resolution volume is divided into M sub-volumes. Each sub-volume is associated to a sequence of N_w c-dimensional volume tokens $\boldsymbol{w} \in \mathbb{R}^{(M \cdot N_w) \times c}$. A 3D patch $\boldsymbol{p}$ in input of the model is divided into N windows. Each window is associated to a sequence of N_v window tokens $\boldsymbol{v} \in \mathbb{R}^{(N \cdot N_v) \times c}$. A window is composed of N_u visual tokens $\boldsymbol{u} \in \mathbb{R}^{(N \cdot N_u) \times c}$ which are the finest subdivision level.

In Fig. 2, volume, window and visual tokens are indicated in green, red, and blue respectively. First, Multi-head Self-Attention (MSA) is performed for each window over the merged sequence of visual and window tokens. Given a sequence of visual tokens *ie.* small patches, MSA is a combination of non-local mean for all the tokens in the sequence [26]. This local operation is denoted as Window-MSA (W-MSA). Second, MSA is performed over the merged sequence of all window tokens and corresponding volume tokens to grasp long-range dependencies in the input patch. This operation is denoted as Global-MSA (G-MSA) and involve only the volume token corresponding to sub-volumes intersecting with $\boldsymbol{p}$. Finally, full resolution attention is achieved by applying MSA over the sequence of volume tokens. Formally, the t-th FINE-transformer bloc is composed of the following three operations:

$$\begin{aligned} [\boldsymbol{u}^t, \hat{\boldsymbol{v}}^t] &= \text{W-MSA}([\boldsymbol{u}^{t-1}, \boldsymbol{v}^{t-1}]), \\ [\boldsymbol{v}^t, \hat{\boldsymbol{w}}^t_{\cap}] &= \text{G-MSA}([\hat{\boldsymbol{v}}^{t-1}, \boldsymbol{w}^{t-1}_{\cap}]), \\ \boldsymbol{w}^t &= \text{MSA}(\boldsymbol{w}^t). \end{aligned} \tag{1}$$

$\boldsymbol{w}_\cap$ denotes the volume tokens corresponding to sub-volumes with a non null intersection with $\boldsymbol{p}$ and $[\boldsymbol{x}, \boldsymbol{y}]$ stands for the concatenation of $\boldsymbol{x}$ and $\boldsymbol{y}$ along the first dimension.

2.3 FINE Properties

Full range interactions. After two FINE transformer blocs, the memory tokens manage to capture global context in the entire volume. This global context can then be propagated to visual tokens from the current patch - see supplementary and Fig. 4.

Complexity MSA has quadratic scaling with respect to the 3D patch dimensions while W-MSA complexity is linear with respect to the input size [16]. FINE only adds a few memory tokens and its complexity is given by:

$$\begin{aligned}\Omega(\text{FINE}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})) =& \Omega(\text{W-MSA}(\boldsymbol{u}, \boldsymbol{v})) + \Omega(\text{G-MSA}(\boldsymbol{v}, \boldsymbol{w}_\cap)) + \Omega(\text{MSA}(\boldsymbol{w})) \\ =& 2c\big(N(N_u + 2N_v) + N_{w_\cap} + MN_w\big)\big(2c + 1\big). \end{aligned} \tag{2}$$

$N_{w_\cap}$ is the number of sub-volumes intersecting with the input patch and can not exceed 8. Only a small number of global tokens brings consistent improvements and we keep $N_v = N_w = 1$. In these conditions, memory tokens are particularly efficient with a negligible complexity overhead compared to W-MSA.

3 Experiments

The Synapse Multi-Atlas Labeling Beyond the Cranial Vault (BCV) [13] dataset is used to compare performances. This dataset comprises 30 CT abdominal images with 7 manually segmented organs per image as ground truth. The organs are spleen (Sp), kidneys (Ki), gallbladder (Gb), liver (Li), stomach (St), aorta (Ao) and pancreas (Pa). The baselines are classic convolutional methods in medical image segmentation [24, 19, 18, 9] and recent state-of-the-art transformer networks [2, 3, 7, 31, 33].

3.1 Data preparation and FINE implementation

All images are resampled to a same voxel spacing. The CT volumes in BCV are not centered, with strong variation along the z (cranio-caudal)-axis. To deal with this issue, the memory tokens are constant along this direction. The sub-volumes are thus reshaped with the same depth as the original volume. FINE is implemented in Pytorch and trained using a single NVidia Tesla V100-32GB GPU. All training parameters (learning rate, number of epochs, data augmentations are provided in the supplementary material). Each training epoch has 250 iterations where a randomly cropped region of size $128 \times 128 \times 64$ voxels is processed. The loss function combines multi-label Dice and cross-entropy losses, and it is optimized using SGD with a polynomial learning rate decay strategy. Deep-supervision is used during training, where the output at each decoder stage is

used to predict a downsampled segmentation mask. To avoid random noise perturbation coming from unseen memory tokens during training (typically memory tokens from regions that have never been selected), a smooth warm-up of these tokens is used. This warm-up consists of masking unseen tokens such that they do not impact the attention or the gradient.

Method	Average		Per organ dice score (%)						
	HD95	DSC	Sp	Ki	Gb	Li	St	Ao	Pa
UNet [24]	-	77.4	86.7	73.2	69.7	93.4	75.6	89.1	54.0
AttUNet [19]	-	78.3	87.3	74.6	68.9	93.6	75.8	89.6	58.0
VNet [18]	-	67.4	80.6	78.9	51.9	87.8	57.0	75.3	40.0
Swin-UNet [2]	21.6	78.8	90.7	81.4	66.5	94.3	76.6	85.5	56.6
nnUNet [9]	10.5	87.0	91.9	86.9	**71.8**	**97.2**	85.3	**93.0**	**83.0**
TransUNet [3]	31.7	84.3	88.8	84.9	72.0	95.5	84.2	90.7	74.0
UNETR [7]	23.0	78.8	87.8	85.2	60.6	94.5	74.0	90.0	59.2
CoTr* [31]	11.1	85.7	93.4	86.7	66.8	96.6	83.0	92.6	80.6
nnFormer [33]	9.9	86.6	90.5	86.4	70.2	96.8	86.8	92.0	83.3
FINE*	**9.2**	**87.1**	**95.5**	**87.4**	66.5	97.0	**89.5**	91.3	82.5

Table 1. Method comparison using the BCV dataset and the training / test split from [33]. Average Dice scores are shown (DSC in % - higher is better). The average and individual organ 95% Hausdorff distances are also shown (HD95 in mm - lower is better). * denotes results trained by us using the authors' public code.

3.2 Comparisons with state-of-the-art

Single fold comparison To fairly compare with reported SOTA results, the same single split of 18 training and 12 test images was used as detailed in [33]. The results are provided in Table 1. FINE obtains the highest average Dice score of 87.1%, which is superior to all other baselines. It also attains the best average 95% Hausdorff distances (HD95) of 9.2mm. Note that the second best method in Dice (nnUNet) is largely below FINE in HD95 (10.5), and the the second best method in HD95 (nnFormer) has a large drop in Dice (86.6).

Method	Average	Sp	Ki	Gb	Li	St	Ao	Pa
CoTr [31]	84.4 ± 3.7	91.8 ± 5.0	87.9 ± 3.4	60.4 ± 10.0	95.7 ± 1.4	84.8 ± 1.3	$\mathbf{90.3} \pm 1.8$	80.0 ± 3.2
nnFormer [33]	84.6 ± 3.6	90.5 ± 6.1	87.9 ± 3.3	63.3 ± 8.1	95.7 ± 1.7	86.4 ± 0.8	89.1 ± 2.0	79.5 ± 3.5
FINE	$\mathbf{86.3} \pm 3.0$	$\mathbf{94.4} \pm 1.9$	$\mathbf{90.5} \pm 4.3$	$\mathbf{65.9} \pm 7.8$	$\mathbf{96.0} \pm 1.1$	$\mathbf{87.9} \pm 1.2$	89.4 ± 1.7	$\mathbf{80.2} \pm 2.8$
P-values	FINE vs. Cotr : 3e-2				FINE vs. nnFormer : 5e-2			

Table 2. Method comparison with SOTA transformer baselines (CoTr and nnFormer) using the BCV dataset and 5-fold cross validation. Results show mean and standard deviation of Dice (in %) for each organ and the average Dice over all organs (higher is better).

5-fold cross-validation comparison 5-fold cross-validation of 18 training and 12 test images was used to compare FINE with the public implementation of the leading transformer baselines (CoTr and nnFormer). The Dice score results are provided in Table 2. FINE's average improvement is significant (more than 1.5 pt with the second baseline with low variance), and FINE gives the best results in 6 out of 7 organ segmentation. The statistical significance in Dice is measured with a paired 2-tailed t-test. The significance of FINE gains with respect to CoTr (3e-2) and nnFormer (5e-2) is confirmed.

3.3 FINE Analysis

Method	WT	VT	Average		Per organ dice score						
			HD95	DSC	Sp	Ki	Gb	Li	St	Ao	Pa
nnFormer [33]	0	0	8.0	86.2	96.0	94.2	57.2	96.5	87.2	89.5	82.5
FINE	✓	0	7.7	86.6	95.7	94.2	60.9	**96.8**	85.1	90.0	**83.8**
	✓	✓	**5.2**	**87.1**	**96.2**	**94.5**	**61.5**	**96.8**	**87.3**	**90.3**	83.0

Table 3. Ablation study of the impact of different tokens on BCV dataset. The metrics are Dice score (DSC in %) for all organs and in average, and the 95% Hausdorff distance (HD95 in mm). WT: Window tokens. VT: Volume tokens.

Method	Memory (GB)
nnUNet [9]	8.6
CoTr [31]	7.62
nnFormer [33]	7.73
FINE	8.05

Table 4. Models memory consumption during training.

Ablation study To show the impact of the different tokens in FINE, an ablation study is presented in Table 3. Three variations of FINE are compared: FINE without tokens, which is equivalent to the nnFormer method; FINE with window tokens but without volume tokens, and FINE with window and volume tokens (default). The results shows that the window tokens generally help to better segment small and difficult organs like the pancreas (Pa) and gallbladder (Gb). The use of window tokens leads to an increase in average Dice by +0.4 points. Furthermore, adding volume tokens increases performance further (average Dice increase of +0.5 points, and average HD95 reduction from 7.7mm to 5.2mm).

FINE complexity The memory consumption of FINE compared to baselines is shown in Table 4. FINE has very low overhead compared to CoTr and nnFormer. In addition, FINE has even a lower consumption than nnUNet.

Visualizations Visualizations of segmentation results from FINE compared to CoTr and nnFormer are presented in Figure 3. All models produce compelling results compared to the ground truth, but one can clearly see differences and especially an improved segmentation from FINE of the spleen. Visualizations of FINE attention maps are provided in figure 4. These attention maps show that FINE is able to leverage context information from the complete image. The left

example shows that there is attention with organs and tissues outside of the crop region (blue rectangle). Furthermore, the right example shows attention from different borders and bones like the spine, which give a strong positional information to the model.

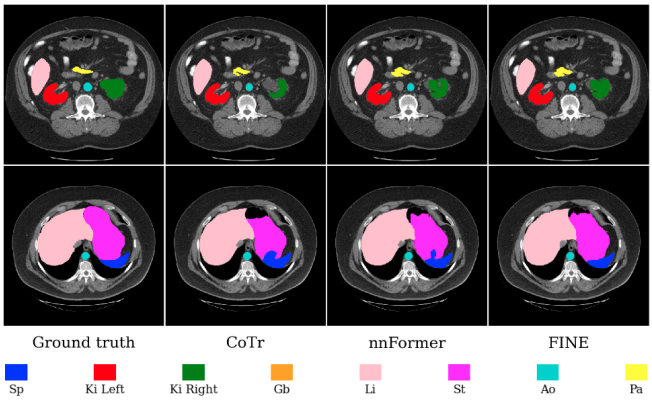


Fig. 3. Visualisation of organs segmentation by FINE compared to state-of-the-art methods on BCV. We can qualitatively see how the full context and high-resolution in FINE help in performing accurate segmentation.

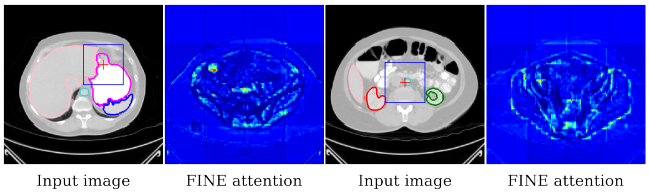


Fig. 4. Visualisation of attention maps of FINE on a BCV segmentation example. The blue rectangle is the sub-volume for which the attention has been calculated.

4 Conclusion

We have presented FINE: the first transformer architecture that allows all available contextual information to be used for automatic segmentation of high-resolution 3D medical images. The technique, using two levels of memory tokens (window and volume), is applicable for any transformer architecture. Results show that FINE improves over recent and state-of-the-art transformers models. Our future work will involve the study of FINE in other modalities such as MRI or US images, as well as for other medical image tasks.

References

1. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation (2021)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation (2021)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
6. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. arXiv preprint arXiv:2104.11227 (2021)
7. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: Unetr: Transformers for 3d medical image segmentation (2021)
8. Hwang, S., Heo, M., Oh, S.W., Kim, S.J.: Video instance segmentation using inter-frame communication transformers (2021)
9. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods (2020)
10. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical Image Analysis **36**, 61–78 (2017). https://doi.org/https://doi.org/10.1016/j.media.2016.10.004, https://www.sciencedirect.com/science/article/pii/S1361841516301839
11. Karimi, D., Vasylechko, S.D., Gholipour, A.: Convolution-free medical image segmentation using transformers (2021)
12. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention (2020)
13. Landman, X., Igelsias, S., Langeral, K.: Multi-atlas labeling beyond the cranial vault. MICCAI (2015)
14. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: Proceedings of the 36th International Conference on Machine Learning. pp. 3744–3753 (2019)
15. Li, H., Yang, F., Zhao, Y., Xing, X., Zhang, J., Gao, M., Huang, J., Wang, L., Yao, J.: Dt-mil: Deformable transformer for multi-instance learning on histopathological image pp. 206–216 (2021). https://doi.org/10.1007/978-3-030-87237-3_20
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. International Conference on Computer Vision (ICCV) (2021)
17. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Proceedings of the 30th International

Conference on Neural Information Processing Systems. p. 4905–4913. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)
18. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation (2016)
19. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas (2018)
20. Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N., Kong, L.: Random feature attention. In: International Conference on Learning Representations (2020)
21. Qiu, J., Ma, H., Levy, O., tau Yih, S.W., Wang, S., Tang, J.: Blockwise self-attention for long document understanding (2020)
22. Rae, J.W., Potapenko, A., Jayakumar, S.M., Hillier, C., Lillicrap, T.P.: Compressive transformers for long-range sequence modelling. In: International Conference on Learning Representations (2019)
23. Reynaud, H., Vlontzos, A., Hou, B., Beqiri, A., Leeson, P., Kainz, B.: Ultrasound video transformers for cardiac ejection fraction estimation pp. 495–505 (2021). https://doi.org/10.1007/978-3-030-87231-1_48
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
25. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation pp. 36–46 (2021). https://doi.org/10.1007/978-3-030-87193-2_4
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
27. Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J.: Boundary-aware transformers for skin lesion segmentation pp. 206–216 (2021). https://doi.org/10.1007/978-3-030-87193-2_20
28. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. arXiv e-prints pp. arXiv–2006 (2020)
29. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvtv2: Improved baselines with pyramid vision transformer. arXiv preprint arXiv:2106.13797 (2021)
30. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: IEEE ICCV (2021)
31. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation (2021)
32. Zhang, P., Dai, X., Yang, J., Xiao, B., Yuan, L., Zhang, L., Gao, J.: Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. ICCV 2021 (2021)
33. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation (2021)