WaveFormer: A 3D Transformer with Wavelet-Driven Feature Representation for Efficient Medical Image Segmentation

Md Mahfuz Al Hasan[1*][0000−0002−7788−388X] (✉), Mahdi Zaman[2*][0000−0003−1668−338X], Abdul Jawad[3*][0000−0001−6530−1451], Alberto Santamaria-Pang[4*][0000−0003−4012−8394], Ho Hin Lee[4][0000−0002−7378−2379], Ivan Tarapov[4][0009−0008−9971−9790], Kyle B. See[1][0000−0003−4180−0963], Md Shah Imran[1][0009−0005−7418−7247], Antika Roy[1][0000−0001−7954−5588], Yaser Pourmohammadi Fallah[2][0000−0002−4920−7104], Navid Asadizanjani[1][0000−0003−3347−5072], and Reza Forghani[1,5][0000−0002−8572−1864]

[1] University of Florida
[2] University of Central Florida
[3] University of California Santa Cruz
[4] Microsoft Healthcare AI
[5] AdventHealth
mdmahfuzalhasan@ufl.edu

Abstract. Transformer-based architectures have advanced medical image analysis by effectively modeling long-range dependencies, yet they often struggle in 3D settings due to substantial memory overhead and insufficient capture of fine-grained local features. We address these limitations with WaveFormer, a novel 3D-transformer that: i) leverages the fundamental frequency-domain properties of features for contextual representation, and ii) is inspired by the top-down mechanism of the human visual recognition system, making it a biologically motivated architecture. By employing discrete wavelet transformations (DWT) at multiple scales, WaveFormer preserves both global context and high-frequency details while replacing heavy upsampling layers with efficient wavelet-based summarization and reconstruction. This significantly reduces the number of parameters, which is critical for real-world deployment where computational resources and training times are constrained. Furthermore, the model is generic and easily adaptable to diverse applications. Evaluations on BraTS2023, FLARE2021, and KiTS2023 demonstrate performance on par with state-of-the-art methods while offering substantially lower computational complexity The code for WaveFormer is publicly available at: https://github.com/mahfuzalhasan/WaveFormer. .

Keywords: Deep Learning · Transformer Network · Discrete Wavelet Transform

* Equal contribution.

1 Introduction

Medical image segmentation is fundamental to clinical applications such as tumor delineation, organ localization, and surgical planning. Deep learning-based approaches, particularly convolutional neural networks (CNNs), have demonstrated significant success by hierarchically extracting features. However, their limited receptive fields hinder the capture of long-range dependencies, a critical shortcoming in 3D applications where spatial context across distant slices is essential. Vision transformers (ViTs) overcome this limitation by employing self-attention to model global dependencies; yet, their application to 3D volumes is often constrained by substantial memory overhead and computational inefficiency. Hierarchical transformers partially address these issues by restricting self-attention to local windows, but they struggle to capture the fine-grained details necessary for precise volumetric segmentation [20,2].

Hybrid CNN-transformer architectures have thus gained popularity by combining the strengths of both paradigms to preserve local detail through CNN modules while harnessing transformers for global reasoning [22,4,24,9]. Although these models enhance performance, they often rely on bulky encoders or complex attention-based decoders—such as UNETR-style architectures [5,7], multi-stage pyramids [1], or large transformer blocks [9]—which lead to excessive parameter counts and slower inference times. These pitfalls make current approaches impractical for deployment in resource-constrained clinical environments, where efficiency and scalability are paramount.

Moreover, neuroscientific evidence suggests that human visual processing involves not only a bottom-up mechanism but also a top-down pathway [6], where coarse, low-spatial frequency information rapidly reaches higher cortical areas to form an abstract representation before merging with local details [3,19]. Motivated by these biologically plausible insights and the need for more efficient transformer designs in 3D, we propose **WaveFormer**. Our approach leverages the frequency-domain properties of volumetric data by applying the discrete wavelet transform (DWT) [18,17] to partition feature maps into low-frequency (global) and high-frequency (local detail) sub-bands. In doing so, WaveFormer explicitly addresses two critical challenges in 3D segmentation: (1) reducing the heavy computational load of global attention and (2) preserving the detailed structures vital for accurate boundary delineation.

Our contributions are threefold: **1. Frequency-Domain Representation Learning:** We compute the bulk of self-attention on the low-frequency subbands, significantly reducing the token count while retaining global context, with parallel streams preserving the high-frequency details essential for accurate segmentation. **2. Efficient Frequency-Guided Decoder:** Instead of relying on conventional upsampling, we adopt an inverse DWT (IDWT) mechanism to reconstruct segmentation masks from high-frequency components. This approach

[6] In the top-down mechanism, low-spatial frequencies are rapidly projected onto the prefrontal cortex to form abstract object representation, which is then back-projected into the temporal cortex for integration with the bottom-up pathway. [3,19]

not only reduces parameter overhead but also enables real-time volumetric inference. **3. Enhanced Local-Global Context Aggregation:** By integrating frequency-domain cues at multiple scales, WaveFormer emulates the top-down processing route of the human visual cortex, effectively fusing coarse global representations with local feature streams to improve 3D segmentation performance.

We evaluate WaveFormer on three major 3D medical benchmarks—BraTS2023, FLARE2021, and KiTS2023. Our experiments demonstrate that WaveFormer achieves competitive or superior accuracy relative to current state-of-the-art methods [22,4,24,9], while substantially lowering model complexity and inference times.

2 WaveFormer

WaveFormer is a hierarchical transformer-based framework designed to address the dual challenges of efficient global context modeling and fine-grained feature preservation while reducing the number of network parameters. As illustrated in Figure 1, it integrates two central design principles:

Efficient Global Context Modeling: WaveFormer reduces computational overhead by applying a discrete wavelet transform (DWT) to extract low-frequency sub-bands, enabling self-attention on a compact representation while preserving essential contextual information.

Detail-Preserving Reconstruction: High-resolution segmentation masks are progressively reconstructed using IDWT, which reintegrates high-frequency sub-bands to recover fine-grained structural details at each decoding stage.

In addition to these wavelet-based operations, a squeeze-and-excitation module [10] serves as a bottleneck for channel calibration, selectively enhancing the most relevant feature channels before the final segmentation output. This mechanism further refines the network's representational capacity without incurring a substantial computational burden.

2.1 Learning on Compact Representations

Recent evidence suggests that self-attention in large-scale transformer models primarily targets low-frequency components [21], acting in effect as a low-pass filter. WaveFormer capitalizes on this by employing the discrete wavelet transform (DWT) to decompose volumetric features into multiple resolution sub-bands, capturing coarse global structures in the low-frequency approximation while preserving crucial fine details in the high-frequency components. By structuring features into discrete frequency bands, the model can attend on a computationally compact low-frequency representation—reducing overhead—yet retains high-frequency information essential for precise boundary delineation.

Unlike traditional decoder-heavy architectures that rely on numerous learnable parameters to progressively recover spatial detail, WaveFormer leverages IDWT for upsampling. This approach reintegrates high-frequency components

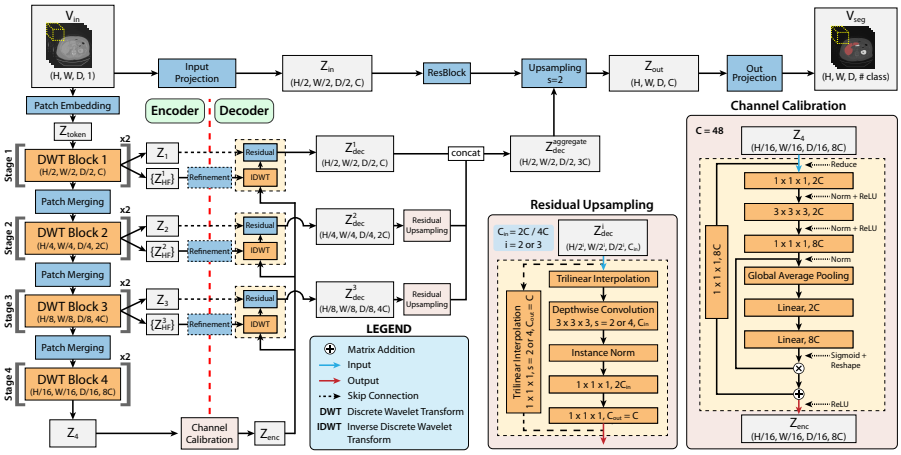

Fig. 1. The overall architecture of the proposed WaveFormer. The block details are provided in Figure 2.

back into the decoded representation, ensuring that the final segmentation output retains both coarse global context and detailed local structures. The multi-resolution capability of wavelets thus offers two key advantages: (i) fewer tokens for global attention, which reduces computational cost, and (ii) efficient reconstruction of fine-scale features, minimizing the need for parameter-intensive decoder blocks. This frequency-centered paradigm opens new possibilities for optimizing transformer-based models, particularly in 3D medical imaging, where multi-scale information can be selectively emphasized to enhance performance and efficiency.

2.2 Encoder

Given a labeled set of 3D images $I_{3D} = \{(X_i, Y_i)\}_{i=1}^{N}$, we randomly crop sub-volumes $V_i \in \mathbb{R}^{H \times W \times D \times P}$ and feed them into the encoder (e.g., $H = W = D = 96$ for FLARE). A simple convolution-based patch embedding reduces the input resolution by a factor of two in each spatial dimension, producing initial tokens $Z_{\text{token}} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times C}$ with $C = 48$. These tokens pass through four sequential encoder stages, each comprising two wavelet-attention blocks (Figure 1).

Wavelet-Attention Block. In each block, the token feature map undergoes a multi-level discrete wavelet transform (DWT) to separate low-frequency (LF) approximation coefficients z_{LF}^{l} from high-frequency (HF) detail coefficients $\{z_{HF}^{l}\}$. As shown in Eq. (1), self-attention (MSA) is computed only on the LF approximation, reducing the computational burden while preserving essential global context. The attention output is then upsampled to match the original spatial resolution of the input tokens to that block, processed through a feed-forward network, and forwarded to the next layer as depicted in Figure 2(a). Symboli-

cally, the wavelet-attention operation in two consecutive layers l and $l+1$ can be written as:

$$
\begin{aligned}
&\textbf{l}^{\text{th}}\ \textbf{Layer} && \textbf{l}+\textbf{1}^{\text{th}}\ \textbf{Layer}\\
&z^{l-1}_{LF_m}, \{z^{l-1}_{HF}\} = \text{DWT}(z^{l-1}, \text{level} = m), && z^{l}_{LF_m}, \{z^{l}_{HF}\} = \text{DWT}(z^{l}, \text{level} = m),\\
&\hat{z}^{l}_{LF_m} = \text{MSA}(\text{LN}(z^{l-1}_{LF_m})), && \hat{z}^{l+1}_{LF_m} = \text{MSA}(\text{LN}(z^{l}_{LF})),\\
&\hat{z}^{l}_{\text{attn}} = \text{Upsample}(\hat{z}^{l}_{LF_m}) + z^{l-1}, && \hat{z}^{l+1}_{\text{attn}} = \text{Upsample}(\hat{z}^{l+1}_{LF_m}) + z^{l},\\
&z^{l} = \text{MLP}(\text{LN}(\hat{z}^{l}_{\text{attn}})) + \hat{z}^{l}_{\text{attn}}, && z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1}_{\text{attn}})) + \hat{z}^{l+1}_{\text{attn}},
\end{aligned}
\tag{1}
$$

where $z^{l}_{LF_m}$ and $\{z^{l}_{HF}\}$ are the LF and set of HF components from an m-level DWT in the l-th layer. $\hat{z}^{l}_{\text{attn}}$ and z^{l} represent the attention output on the LF approximation and final output from the l-th layer respectively. z^{l} goes through a similar computational flow in the next attention layer of a block and provides the final output from the encoder stage i $z^{l+1} = z^{i}$. HF component set from both layer ($\{z^{l}_{HF}\}$ and $\{z^{l+1}_{HF}\}$) are merged to produce final fused HF components from encoder stage i $\{z^{i}_{HF}\}$.

Hierarchical Encoding. Following the strategy in [8], each encoder stage concludes with a patch merging operation that downsamples the feature map by a further factor of two and increases the channel dimension accordingly. Thus, Stage 1 operates on tokens of size $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times C$, Stage 2 on $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 2C$, Stage 3 on $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 4C$, and Stage 4 on $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 8C$. Each stage's output tokens and HF details $\{z^{l}_{HF}\}$ are relayed to the decoder. In Stage 4, DWT is omitted because the tokens are already at the lowest spatial resolution.

The network is designed so that, within each wavelet-attention block, the features undergo multi-level DWT with decreasing decomposition levels m at progressively deeper stages. For instance, Stage 1 applies DWT with $m = 3$ on tokens of resolution $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$ (i.e., down to $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$) before performing attention, while Stages 2 and 3 use $m = 2$ and $m = 1$, respectively. This progressive reduction preserves hierarchical representations of global context while controlling computational complexity.

Multiscale Attention. Utilizing the multi-scale approximation feature from DWT in each layer, we extended our network to compute multi-resolution attention as depicted in Figure 2(b). For stage 1, attention is computed on each DWT decomposed feature on $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$ and $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$ scales, respectively. Window attention with the window size matching the lowest-scale feature resolution (3^{rd} level in stage 1) enables capturing contextual relations in both local and global regions within each attention layer, leading to more holistic representation learning. The procedure similarly continues in the 2^{nd} and 3^{rd} stage encoder blocks with 2 scales and 1 scale attention computations, respectively.

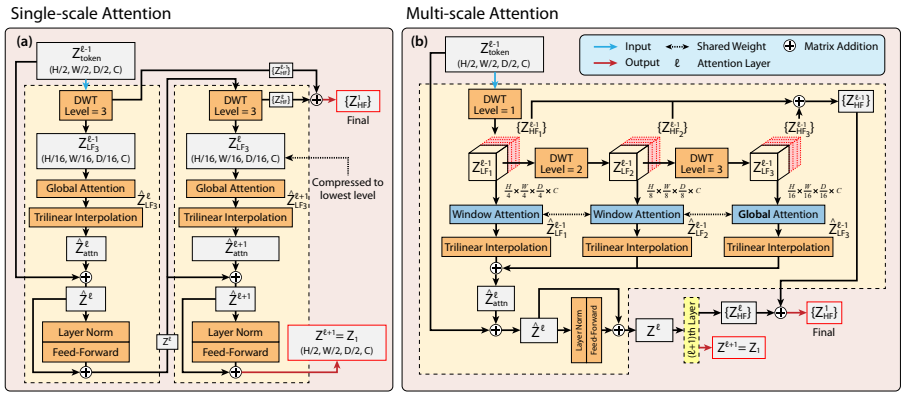

Fig. 2. Block architecture of the proposed network for encoder stage 1. (a) Single-scale Attention: Attention is computed solely on the final approximation coefficients (Z_{LF_3}) obtained from multi-level (3 in stage 1) DWT on input token. The HF components extracted at both attention layers are combined (Z^1_{HF}) and passed to the decoder along with the output from the current stage (Z_1) and final encoder output (see Figure 1). (b) Multi-scale Attention: Attention is computed at each resolution level of the decomposed token obtained by the DWT. A fixed window size matching the lowest resolution (DWT level 3) enables global attention on DWT level 3 while local window attention on other decomposition levels (1 & 2), leading to the capture of both global and local context in a single attention layer.

2.3 Decoder

At the end of the encoder, the hidden dimension feature map z_4 is calibrated using a squeeze-and-excitation module [10] to produce a refined embedding z_{enc}. z_{enc} is subsequently input into an IDWT upsampling path along with intermediate encoder features z_i and the associated high-frequency (HF) coefficients $\{z^i_{HF}\}$ where $i = 1, 2, 3$. A lightweight refinement block suppresses noise in the high-frequency details, and the IDWT reconstructs an upsampled representation that merges with the corresponding skip connection z_i to generate decoder feature map z^i_{dec} of dimensions $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 4C$, $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 2C$, $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times C$ at stages 3, 2, and 1, respectively. The outputs from stages 3 and 2 are further upsampled using a residual module [16] and concatenated with the Stage 1 output to form an aggregated feature $z^{\text{aggregate}}_{\text{dec}}$. Finally, fusing $z^{\text{aggregate}}_{\text{dec}}$ with the original patch embedding x_{in} and applying a projection layer produces the final decoded tensor $Z_{\text{out}} \in \mathbb{R}^{H \times W \times D \times C}$ and the segmentation mask $V_{\text{pred}} \in \mathbb{R}^{H \times W \times D \times \text{class}}$.

3 Experiments

3.1 Datasets

We assessed WaveFormer along with comparable baseline models on three tasks utilizing three publicly available datasets: FLARE2021 [13] contains 361 ab-

domen CT volumes for multi-organ segmentation. Each volume includes four segmentation targets (spleen, kidney, liver, and pancreas). KiTS2023 [15] contains 489 abdomen CT volumes (publicly released version), featuring three segmentation targets (kidney, renal tumors and renal cyst). BraTS2023 [11,14] contains a total of $1,251$ 3D brain MRI volumes, each including four modalities (namely T1, T1Gd, T2, T2-FLAIR) and three segmentation targets (WT: Whole tumor, ET: Enhancing tumor, TC: Tumor core).

3.2 Implementation Details

We use PyTorch 2.3.1-CUDA12.1 and Monai 0.9.0 to implement our experimental framework as well as the baseline models. For **FLARE2021** and **KiTS2023**, our training scheme closely follows [12], using volume patches of size $96 \times 96 \times 96$. Each input is resampled to $1.0 \times 1.0 \times 1.2mm^3$ spacing and the models are trained for $40,000$ iterations. For **BraTS2023**, we closely follow the training scheme provided in [23]. Input samples are randomly cropped to $128 \times 128 \times 128$, and models are trained for $250K$ iterations. For all training, We use an AdamW optimizer with a learning rate of 0.0001, Dice and CE loss as objective functions, and report the Dice similarity coefficient (Dice score) along with 95^{th} percentile Hausdorff Distance (HD95) for assessing volumetric accuracy. Mean scores from 5-fold cross-validation with an 80:20 split are reported for **FLARE** and **KiTS**. During BraTS training, the test set provided by the authors of [23] is used for final evaluation, while the remaining dataset is split into an 80:20 ratio. All experiments used a batch size of 2 per GPU, with 1 A100 GPU per training for FLARE and KiTS and 4 A100 GPU for BraTS.

3.3 Comparison with SOTA Methods

The segmentation results for the BraTS2023 dataset are listed in Table 1. On **BraTS2023**, our model WaveFormer achieves the highest overall 91.37% mean Dice score, with scores of 93.71% and 88.47% on WT and ET, respectively, and a better HD95 for TC and ET (see Table 1). On FLARE2021, we achieved a superior overall and organ-wise (spleen, liver, and pancreas) Dice score segmentation with significantly fewer parameters compared to state-of-the-art transformer and hybrid models as shown in Table 2. On KiTS2023, our model shows better overall mean Dice score (background included). KiTS2023 is a large dataset with extremely high-resolution CT scans. Due to resource constraints, we could not run different variants of our network. We plan to provide detailed results on KiTS in future work.

3.4 Ablation of Architectural Improvements

We evaluate the impact of key architectural components on tumor segmentation across varying tumor sizes (Table 3). While HF refinement slightly improves

Table 1. Quantitative comparison on the BraTS2023 dataset. Best scores in bold. SegMamba parameter count was unavailable from public sources.

Methods	Params↓	BraTS2023							
		WT		TC		ET		Avg	
		Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
UX-Net [12]	53M	93.13	4.56	90.03	5.68	85.91	4.19	89.69	4.81
MedNeXt [16]	18M	92.41	4.98	87.75	4.67	83.96	4.51	88.04	4.72
UNETR [7]	92.8M	92.19	6.17	86.39	5.29	84.48	5.03	87.68	5.49
SwinUNETR [6]	62.2M	92.71	5.22	87.79	4.42	84.21	4.48	88.23	4.70
SwinUNETR-V2 [8]	72.8M	93.35	5.01	89.65	4.41	85.17	4.41	89.39	4.51
SegMamba [23]	–	93.61	**3.37**	**92.65**	3.85	87.71	3.48	91.32	3.56
Our Method	**16.97M**	**93.71**	3.64	91.94	**3.67**	**88.47**	**3.26**	**91.37**	**3.52**

Table 2. Dice score comparison on the FLARE2021 and KiTS2023 datasets. Evaluated with background inclusion. Best scores in bold.

Methods	Params↓	FLARE2021					KiTS2023
		Spleen	Kidney	Liver	Pancreas	Mean↑	Mean↑
TransBTS [22]	31.6M	96.4	95.9	97.4	71.1	90.2	75.56
UNETR [7]	92.8M	92.7	94.7	96.0	71.0	88.6	70.23
SwinUNETR [6]	62.2M	97.9	96.5	98.0	78.8	92.9	80.74
nnFormer [24]	149.3M	96.0	**97.5**	97.7	71.7	90.8	78.51
3D-UXNET [12]	53M	98.1	96.9	**98.2**	80.1	93.4	78.88
Ours	**16.9M**	**98.2**	97.0	**98.2**	**81.7**	**93.8**	**80.91**

small tumor segmentation, it adversely affects medium and large tumors, reducing overall performance. In contrast, multi-scale attention with a channel-calibrated bottleneck (without HF refinement) yields the best results, especially for small-sized TC and ET tumors. Decreasing the decomposition level degrades performance, particularly for small tumors. However this also highlights the modular nature of our network. We believe optimizing decomposition levels for task-specific organ segmentation is a promising direction for future research.

4 Conclusion

In this work, we introduce WaveFormer, a novel 3D Transformer architecture leveraging discrete wavelet transforms for efficient medical image segmentation. WaveFormer effectively captures global context and fine-grained details, significantly reducing computational overhead and parameter count. Experimental results demonstrate superior performance compared to state-of-the-art models in specific segmentation tasks, maintaining efficiency without compromising ac-

Table 3. Dice score comparison on BraTS across scales and architectural variations. Binning for Small(S), Medium(M), Large(L) targets (in cm^3): WT: [0–71, 71–120, >120], TC: [0–20, 20–40, >40], ET: [0–12, 12–30, >30]

Methods	Params	Variations	DWT Level	WT (%)			TC (%)			ET (%)			Mean
				S	M	L	S	M	L	S	M	L	
Single-scale Attn	17.06M	with refine.	3,2,1,0	91.44	94.67	95.26	85.88	93.98	94.77	81.51	92.72	92.71	91.12
	16.97M	w/o refine.		91.41	94.58	95.24	85.83	94.75	94.70	81.26	92.72	93.17	91.3
Multi-scale Attn	16.97M	conv bottleneck	3,2,1,0	89.87	94.12	95.17	85.99	94.80	95.15	80.37	92.47	93.07	91.24
	16.97M	Channel Calib.		91.04	94.67	95.40	87.90	94.74	94.96	82.13	92.64	93.07	91.37
	16.97M	Channel Calib.	2,2,1,0	90.36	94.79	95.19	83.83	94.07	95.57	82.69	92.53	93.03	90.92

curacy. Our findings highlight the promise of frequency-domain representations for developing lightweight, effective deep learning solutions in medical imaging.

Acknowledgments. We are thankful to brAIn lab and its associates and National Research Platform for providing necessary computation resources throughout the project.

Disclosure of Interests. The authors declare that they have no conflict of interest.

References

1. Azad, R., Kazerouni, A., Azad, B., Khodapanah Aghdam, E., Velichko, Y., Bagci, U., Merhof, D.: Laplacian-former: Overcoming the limitations of vision transformers in local texture detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 736–746. Springer (2023)
2. Bai, J., Yuan, L., Xia, S.T., Yan, S., Li, Z., Liu, W.: Improving vision transformers by revisiting high-frequency components. In: European Conference on Computer Vision. pp. 1–18. Springer (2022)
3. Bar, M.: A cortical mechanism for triggering top-down facilitation in visual object recognition. Journal of cognitive neuroscience **15**(4), 600–609 (2003)
4. 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. arXiv preprint arXiv:2102.04306 (2021)
5. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2022)
7. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
8. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer (2023)

9. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J., Merhof, D.: 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)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
11. Kazerooni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., et al.: The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). ArXiv (2023)
12. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv:2209.15076 (2022)
13. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al.: Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. Medical Image Analysis **82**, 102616 (2022)
14. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
15. Myronenko, A., Yang, D., He, Y., Xu, D.: Automated 3d segmentation of kidneys and tumors in miccai kits 2023 challenge. In: International Challenge on Kidney and Kidney Tumor Segmentation, pp. 1–7. Springer (2023)
16. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)
17. Santamaria-Pang, A., Qiu, J., Chowdhury, A., Kubricht, J., Tu, P., Naresh, I., Virani, N.: Adversarial attacks with time-scale representations. arXiv preprint arXiv:2107.12473 (2021)
18. Shensa, M.J., et al.: The discrete wavelet transform: wedding the a trous and mallat algorithms. IEEE Transactions on signal processing **40**(10), 2464–2482 (1992)
19. Ullman, S.: Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. Cerebral cortex **5**(1), 1–11 (1995)
20. Wang, P., Zheng, W., Chen, T., Wang, Z.: Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. arXiv preprint arXiv:2203.05962 (2022)
21. Wang, Z., Luo, H., Wang, P., Ding, F., Wang, F., Li, H.: Vtc-lfc: Vision transformer compression with low-frequency components. Advances in Neural Information Processing Systems **35**, 13974–13988 (2022)
22. Wenxuan, W., Chen, C., Meng, D., Hong, Y., Sen, Z., Jiangyun, L.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 109–119 (2021)
23. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15008. Springer Nature Switzerland (October 2024)
24. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)