MetaUNETR: Rethinking Token Mixer Encoding for Efficient Multi-Organ Segmentation

Pengju Lyu[1,2], Jie Zhang[3], Lei Zhang[4], Wenjian Liu[2](✉), Cheng Wang[1], Jianjun Zhu[1,5](✉)

[1] Hanglok-Tech Co., Ltd., Hengqin 519000, China
jj.zhu@hanglok-tech.cn
[2] City University of Macau, Macau 999078, China
andylau@cityu.edu.mo
[3] Guangdong Provincial Key Laboratory of Tumor Interventional Diagnosis and Treatment, Zhuhai People's Hospital, Zhuhai Hospital Affiliated with Jinan University, Zhuhai, Guangdong 519000, China
[4] Laboratory of Vision Engineering, School of Computer, University of Lincoln, LN6 7TS, UK
[5] Center of Interventional Radiology & Vascular Surgery, Department of Radiology, Zhongda Hospital, Medical School, Southeast University, Nanjing 210009, China

Abstract. The Transformer architecture and versatile CNN backbones have led to advanced progress in sequence modeling and dense prediction tasks. A critical development is the incorporation of different token mixing modules such as ConvNeXt, Swin Transformer. However, findings within the MetaFormer framework suggest these token mixers have a lesser influence on representation learning than the architecture itself. Yet, their impact on 3D medical images remains unclear, motivating our investigation into different token mixers (self-attention, convolution, MLP, recurrence, global filter, and Mamba) in 3D medical image segmentation architectures, and further prompting a reevaluation of the backbone architecture's role to achieve the trade off in accuracy and efficiency. In the paper, we propose a unified segmentation architecture—MetaUNETR featuring a novel TriCruci layer that decomposes the token mixing processes along each spatial direction while simultaneously preserving precise positional information on its orthogonal plane. By employing the Centered Kernel Alignment (CKA) analysis on feature learning capabilities among these token mixers, we find that the overall architecture of the model, rather than any specific token mixers, plays a more crucial role in determining the model's performance. Our method is validated across multiple benchmarks varying in size and scale, including the BTCV, AMOS, and AbdomenCT-1K datasets, achieving the top segmentation performance while reducing the model's parameters by about 80% compared to the state-of-the-art method. This study provides insights for future research on the design and optimization of backbone architecture, steering towards more efficient foundational segmentation models. The source code is available at https://github.com/lyupengju/MetaUNETR.

Keywords: Token mixers · Multi-organ segmentation· Architecture optimization.

1 Introduction

Significant strides in the field of deep learning (DL) have been notably propelled by advancements in sequence models [5]. Their extensive versatility is empirically showcased across diverse domains including natural language understanding, time series forecasting as well as computer vision (e.g., ViT [1] when images are cast into sequentially patchified representations). The formulation of DL models for sequential data revolves around the conceptualization of sequence-to-sequence transformations, leveraging basic token mixers such as convolutions, self-attention or recurrence, and consequently giving rise to the prominent families of deep sequence models [2], namely, convolutional neural networks (CNNs), Transformers, recurrent neural networks (RNNs), and recent emergence of state space model notably exemplified by Mamba [3,14].

The transformative impact of deep sequence models has in recent years reshaped the landscape of medical image analysis. Multi-organ segmentation, a critical task in understanding anatomical structures for clinical diagnosis and treatment planning, has witnessed a paradigm shift from CNNs to Transformers which demonstrate efficacy for capturing long-range dependencies [19]. Under the prevailing notion that the attention-based token mixer module is the primary contributor to the competence of Transformers, Since UNETR [4], pioneering works have been focusing on integrating diverse self-attention variants as backbone cores within the U-Net topology [18], for instance, Swin UNETR [20], and UNesT [24] employed Swin Transformer [11] and Nested Transformer [25] respectively, achieving promising performance in modeling high anatomical variability. Subsequent investigations have also revealed that the overarching architecture of Transformers, often referred to as MetaFormer [23], plays a more pivotal role in determining the models' performance. The success achieved by modernized ConvNeXt [12] and MLP-Mixer [22] within the MetaFormer architecture challenges the predominant emphasis on attention, signifies a resurgence of CNN and MLP paradigms [22]. 3D UX-NET [9] substantiates this notion by substituting self-attention modules in Swin UNETR with efficient large kernel depthwise convolutions, yet achieving state-of-the-art multi-organ segmentation accuracy. Despite the notable achievements garnered by employing various novel token mixers encoded Metaformer-style models in multi-organ segmentation, the detailed comprehension of why specific mixers outperform others in this domain remains largely unexplored. Additionally, these models, regardless of the scale of training data set and hardware constraints frequently incur compromised computation efficiency. There is a desire to enhance and optimize the architecture to attain superior segmentation performance, though this has not yet become the primary focus.

In this study, our objective is to understand the learning differences among various mixer types and their influence on overall volumetric segmentation tasks. We aim to answer the following research questions regarding mixer selection and backbone architecture design: 1) Which mixer type has superior and more efficient representational learning capabilities in 3D medical image segmentations? 2) The importance and impact of the architecture in comparison to these token mixers. 3) Whether the backbone architecture can be optimized to achieve the superior segmentation performance balancing the tradeoff between accuracy and efficiency. To this end, we propose MetaUNETR comparing 6 token mixers spanning self-attention (Attn) [20], convolution (Conv) [12], MLP [22], recurrence (Recur) [21], global filter (GF) [17], and Mamba [3] for multi-organ segmentation. Our findings indicate that the efficacy of MetaUNETR is primarily derived from the learning capabilities of the top shallow layers, irrespective of the specific type of token mixers utilized for feature learning. Our main contributions are summarized as follows: 1) We propose a generic segmentation network-MetaUNETR featuring a lightweight TriCruci layer, specifically designed for independent token mixing along spatial directions. 2) Leveraging the Centered Kernel Alignment (CKA) analysis, we demonstrate the reduced dominance of various token mixers on feature learning and highlight the importance of features within the upper encoder layers while identifying redundant computations in the deeper layers. 3) Through validation on the BTCV, AMOS, and AbdomenCT-1K datasets, MetaUNETR realizes state-of-the-art segmentation performance with a significantly smaller model size (with 20% parameters) compared to prior arts.

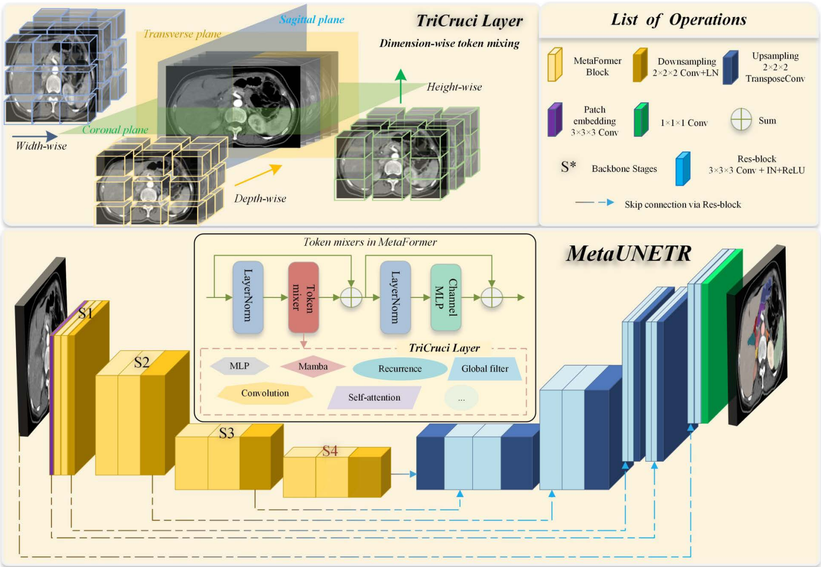

Fig. 1. Architectural overview of the proposed MetaUNETR framework. The MetaFormer encoder backbone integrates diverse token mixer modules, including self-attention, convolution, MLP, recurrence, global filter, and Mamba. The TriCruci layer performs efficient token mixing decomposed along each spatial dimension (depth, height and width) before fusion, synergistically encoding three-dimensional contextual information.

2 Methods

2.1 MetaUNETR architecture

Our MetaUNETR architecture adopts a hierarchical multi-scale encoder-decoder design inspired by UNETR [4] to effectively model both local details and global context for volumetric segmentation, as seen in Figure 1. The input image, characterized by a sub-volume $\boldsymbol{X} \in R^{H \times W \times D \times C}$ with H, W, D and C denoting the spatial dimensions and the number of channels, is initially partitioned into a sequence of patches (patch size = 2) and projected into an embedding space of dimension $S = 48$ through a stem convolution layer. The resultant embedded tokens $\boldsymbol{X}' \in R^{L \times S}, L = \frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$, are fed to a series of MetaFormer blocks, constituting four encoder stages wherein the channel number doubles while feature maps reduce their resolution by half. Each encoder stage comprises two residual sub-blocks with the first one housing a token mixer layer while the subsequent sub-blocks integrating an inverted bottleneck MLP layer. Through residual blocks-enhanced skip connections, the multi-resolution encoder features are concatenated with the decoder composed of residual blocks and transposed convolutions for fine-grained feature recovery. The ultimate decoder feature is processed through a $1 \times 1 \times 1$ convolutional layer in conjunction with a Softmax layer, culminating in the generation of the final segmentation probability map.

2.2 Token mixers in MetaUNETR

The fundamental purpose of token mixers is generally to enhance the original feature space of input sequence X through introducing an inductive bias characterized as dependencies during a designated 'mixing' process, which is parametrically defined by θ and determining how information across X is aggregated. We categorize token mixers into the following two paradigms: 1) Content-Dependent Mixing. The mixing process is intricately tied to the content of the input sequence, allowing for adaptability and context-aware feature interactions, e.g., self-attention and Mamba. 2) Content-Agnostic Mixing. The mixing process is determined solely by the internal parameters of the token mixer and remains fixed across all input signals, e.g., depthwise convolution, MLP, global filter, and recurrence.

$TokenMixer\left(\cdot|\theta\right) \not\perp\!\!\!\perp X :$

$$\begin{aligned}
&Y_t^{Attn} \ := Soft\max\left(Q\left(X_t\right)K\left(X\right)^T\right)V\left(X\right); \\
&Y_t^{Mamba} := C\left(X_t\right)\left(\bar{A}\left(X_t\right)h_{t-1} + \bar{B}\left(X_t\right)X_t\right) \\
&\bar{A}, \bar{B} \leftarrow discretize\left(\mathit{\Delta}\left(X_t\right), A, B\right);
\end{aligned} \tag{1}$$

$$
\begin{aligned}
&TokenMixer\left(\cdot|\theta\right) \perp\!\!\!\perp X:\\
&\qquad Y_t^{Conv} := \sum_{k=0}^{k-1} X_{-k+t} w_k^{Conv};\\
&\qquad Y_t^{MLP} := \left(X^T W_t^{MLP}\right)^T;\\
&\qquad Y^{GF} := \mathcal{F}^{-1}\left(W^{GF} \odot \mathcal{F}\left(X\right)\right);\\
&\qquad Y_t^{Recur} := W_t^{MLP}\left(\tanh\left(W_{hh}^{MLP} h_{t-1} + W_{hx}^{MLP} X_t\right)\right).
\end{aligned} \tag{2}
$$

For the sake of simplicity, we exclude considerations related to scaling, normalization, and biases. X_t/Y_t denote the t_{th} token of input/output sequence and h_t is implicit latent states. $\mathcal{F}$ represent discrete Fourier transform operator, $W^{tokenmixer}$ signifies learnable static weight while $Q, K, V, \bar{A}, \bar{B}, \varDelta$ are dynamic embeddings from X.

2.3 TriCruci layer

To circumvent the substantial computational overhead and proliferation of model parameters arising from the flattening of volumetric data into sequential representations requisite for token mixing operations, motivated by the effectiveness of multi-branch design [10], we propose a novel Triple-Cruciform (TriCruci) architecture design for integrating three-dimensional spatial knowledge in our token-mixer layer. Conceptually, it synergistically amalgamates orthogonal spatial cues along the cardinal dimensions of height, width, and depth for volumetric data representation learning while preserving the precise positional information on their respective perpendicular planes, i.e., coronal (co), sagittal (se) and transverse (tr), as illustrated in Figure 1.

For an input $\boldsymbol{X} \in \mathbb{R}^{H\times W\times D\times C}$, $\left\{\boldsymbol{X}_{:,tr,:}^{depth} \in \mathbb{R}^{D\times C}\right\}_{tr=1}^{H\times W}$ can be regarded as an ensemble of sequences, where D signifies the quantity of tokens along the depth direction, C denotes the channel dimension and $H \times W$ represent the number of sequences in the transverse plane. Following Equation 1 and 2 , token mixing is applied to all input sequences $\boldsymbol{X}_{:,tr,:}^{depth}$ in a weight-sharing paradigm. We combine: $\left\{\boldsymbol{Y}_{:,tr,:}^{depth} \in \mathbb{R}^{D\times C}\right\}_{tr=1}^{H\times W}$ into: $\boldsymbol{Y}^{tr} \in \mathbb{R}^{H\times W\times D\times C}$,$\left\{\boldsymbol{X}_{:,co,:}^{height} \in \mathbb{R}^{H\times C}\right\}_{co=1}^{W\times D}$ into $\boldsymbol{Y}^{co} \in \mathbb{R}^{H\times W\times D\times C}$ and $\left\{\boldsymbol{Y}_{:,sa,:}^{width} \in \mathbb{R}^{W\times C}\right\}_{sa=1}^{H\times D}$ into $\boldsymbol{Y}^{sa} \in \mathbb{R}^{H\times W\times D\times C}$ in similar manner, which are then summed and processed point-wisely in a fully-connection layer $FC\left(\cdot\right)$:

$$\boldsymbol{Y} = FC\left(Sum\left(\boldsymbol{Y}^{tr}, \boldsymbol{Y}^{co}, \boldsymbol{Y}^{sa}\right)\right). \tag{3}$$

2.4 Measuring Representational similarity with linear CKA

We employ linear Centered Kernel Alignment (CKA) [7,16] to assess the alignment or divergence of learned representations among various neural network architectures. Let $\boldsymbol{X}_1 \in \mathbb{R}^{m\times n_1}$, $\boldsymbol{X}_2 \in \mathbb{R}^{m\times n_2}$ denote network layer activations pertaining to the same set of m examples in a minibatch, $n_1 = n_2 =$

$H \times W \times D \times C$. Ranging from 0 (orthogonal) to 1 (identical), CKA leverage the Hilbert-Schmidt Independence Criterion (HSIC) with a linear kernel to gauge the similarity between the centered kernel matrices $\tilde{\boldsymbol{K}}_1 = \boldsymbol{H}\boldsymbol{K}_1\boldsymbol{H}$ and $\tilde{\boldsymbol{K}}_2 = \boldsymbol{H}\boldsymbol{K}_2\boldsymbol{H}$: $HSIC\left(\tilde{\boldsymbol{K}}_1\tilde{\boldsymbol{K}}_2\right) = vec\left(\tilde{\boldsymbol{K}}_1\right) vec\left(\tilde{\boldsymbol{K}}_2\right) / (m-1)^2$, where the two kernel matrices $\boldsymbol{K}_1 = \boldsymbol{X}_1{\boldsymbol{X}_1}^T$ and $\boldsymbol{K}_2 = \boldsymbol{X}_2{\boldsymbol{X}_2}^T$ encapsulate the pairwise similarities between examples, $\boldsymbol{H} = \boldsymbol{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T$ is the centering matrix with I_m being the identity matrix and 1 being a vector of ones of proper size. CKA finally normalizes HSIC to yield a similarity index immune to isotropic scaling:

$$CKA\left(\boldsymbol{K}_1, \boldsymbol{K}_2\right) = \frac{HSIC\left(\tilde{\boldsymbol{K}}_1\tilde{\boldsymbol{K}}_2\right)}{\sqrt{HSIC\left(\tilde{\boldsymbol{K}}_1\tilde{\boldsymbol{K}}_1\right) HSIC\left(\tilde{\boldsymbol{K}}_2\tilde{\boldsymbol{K}}_2\right)}}. \tag{4}$$

In practice, we obtain CKA as an average HSIC score over minibatches.

3 Experiments

3.1 Datasets and implementation details

To comprehensively evaluate impact of token mixers on model performance in the context of multi-organ segmentation tasks, we deliberately curated a triad of public datasets exhibiting progressive enlargement in population scale -BTCV [8] (small), AMOS [6] (medium), and AbdomenCT-1K [15] (large) datasets. We employ 30 CT scans featuring detailed delineation of 13 distinct organs in BTCV, 300 and 1000 multi-contrast abdominal CT from AMOS and AbdomenCT-1K spanning 16 and 4 anatomies respectively. The data preprocessing procedures applied to these three datasets are consistent with the those outlined in UNesT [24] (BTCV), 3D UX-NET [9] (AMOS), MICCAI FLARE23 Challenge [13] (AbdomenCT-1K).

We implemented MetaUNETR using the PyTorch[1] and MONAI[2] frameworks on 4 NVIDIA RTX3090 GPUs. The AdamW optimizer was utilized against a combination of cross-entropy and soft Dice losses. Model performances are reported in terms of Dice score. Further details regarding dataset profiles and training protocols can be found in the supplementary material.

3.2 Model Comparison among different token mixers

In order to compare the efficacy of MetaUNETR employing diverse token mixing backbones, we evaluated model performances under uniform computational resource constraints characterized by commensurate FLOPs (327.64-330.48G) and number of parameters (68.23-68.94M). In Figure 2, validation Dice scores were

[1] http://pytorch.org/

[2] https://monai.io/

recorded over the course of training across the BTCV, AMOS, and AbdomenCT-1K corpora, which reveals a predominantly comparable segmentation efficacy demonstrated by the 6 token mixer configurations. Specifically, for the BTCV and AMOS cohort, peak Dice averaged over the 10 anatomical targets reached 0.825 and 0.87 respectively with marginal dice fluctuations while AbdomenCT-1K exhibited maximum Dice of 0.945 and superior stability. The broadly competitive accuracies attained irrespective of exact encoder selections imply architectural flexibility in satisfying precision standards under fixed computational budgets.

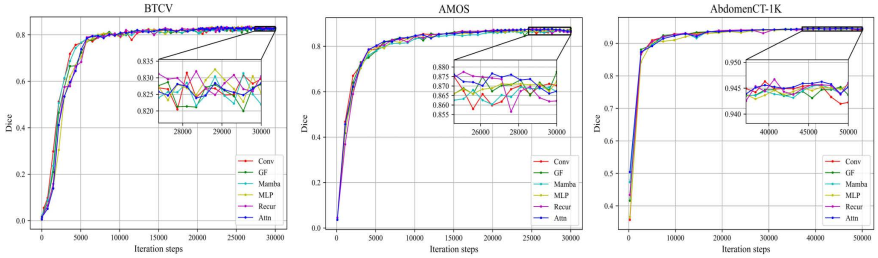

Fig. 2. The comparisons of segmentation performance on validation datasets during training on the MetaUNETR across six token mixers. The achieved proximity Dice scores reveal a reduced dominance of token mixers on feature learning against the architecture.

To further investigate the model learning dynamics across token mixer backbones and unraveling the reason behind their similar segmentation performance, we conducted a comparative evaluation of learned representational similarities utilizing CKA applied to activations following the LayerNorm layer (5 at each stage) of the models. The evaluation was carried out on models trained on the BTCV dataset, as well as 200 CT scans from the AMOS and 300 from the AbdomenCT-1K. The inputs consisted of central image cubes, cropped in accordance with the anatomical regions delineated by segmentation labels to focus comparisons on target areas. Mini-batch size was set to 12 examples during similarity computations. CKA results in Figure 3 reveal that self-attention demonstrate highly concordant early-stage feature representations with alternate token mixer variants, while diverging primarily in stage 4 encodings. This finding suggests that a fourth stage provides limited additional specialization. Analogous conclusions can be drawn from other pairwise similarity comparisons included in the supplementary materials. The encoding similarity patterns suggest flexibility in earlier-stage architectural configurations and depth specifications.

Inspired by previous findings, we eliminated the last stage of MetaUNETR and conducted a comparative analysis against state-of-the-art models based on convolutional (3D UX-NET) and self-attention mechanisms (UNesT, Swin UNETR) across three datasets. As shown is Table 1 MetaUNETR demonstrates

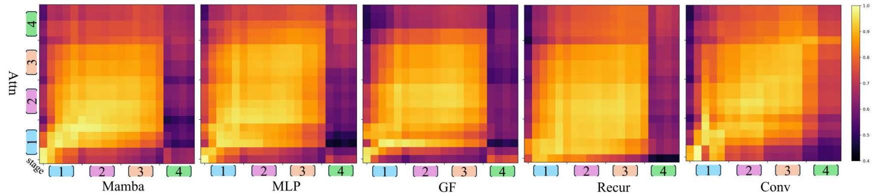

Fig. 3. CKA analysis comparing the representational similarity of intermediate features across different token mixer in MetaUNETR. Substantial similarity is evinced among encodings of the top three stages, while divergences manifest in the final one.

Table 1. Quantitative comparisons for multi-organ segmentation among three-layer MetaUNETR and prior arts. The best results are indicated in bold.

Models		Swin UNETR (CVPR)	3D UX-NET (ICLR)	UNesT (MedIA)	MetaUNETR (3 stages)					
					Attn	Conv	MLP	Recur	GF	Mamba
Dice	BTCV	0.806	0.810	0.813	0.818	0.815	0.813	**0.821**	0.819	0.820
	AMOS	0.887	0.881	0.886	0.893	0.890	0.890	**0.897**	0.895	0.893
	AbdomenCT-1K	0.940	0.938	0.941	0.940	0.938	0.938	**0.942**	0.940	0.940
FLOPs (G)		331.56	639.45	261.73	64.53	63.65	63.52	63.95	**62.87**	63.62
Params (M)		69.94	53.08	87.30	14.70	14.55	14.93	14.95	**14.31**	14.37

comparable performance, achieving approximately 0.940 on the AbdomenCT-1K dataset and marginally outperforming prior arts on the BTCV and AMOS datasets, validating the effectiveness of the TriCruci layer. Notably, this commendable performance is achieved while necessitating substantially fewer computational resources. This efficiency gain not only translates into a more judicious utilization of GPU memory but also facilitates an expansion of the training batch size, thereby contributing to heightened computational efficiency and accuracy. Corresponding visual demonstrations can be found in the supplementary material.

4 Conclusion

In this study, we explored various token mixers and their impact on the MetaFormer architecture for 3D multi-organ segmentation tasks. Our findings indicate that the general architecture of MetaFormers is more vital for model performance than any specific token mixer module. Comparative analyses using CKA uncovered significant similarities and the importance of features within the upper encoder layers while also identifying redundant computation in the bottom layers. Based on these findings, our proposed slim MetaUNETR optimizes the architecture by pruning the redundant encoder stage and includes a novel TriCruci layer for independent mixing processing in each spatial direction, enhancing the mapping of long sequences. Experimental evaluations on the BTCV, AMOS, and

AbdomenCT-1K datasets demonstrated superior performance across all token mixers, with the model size being significantly reduced by approximately 80% compared to state-of-the-art multi-organ segmentation methods. We believe that the MetaUNETR, with its design able to encapsulate a variety of token mixers as modular components, can enhance the flexibility of encoder design and its potential to reduce the architecture search space within AutoML frameworks facilitating the creation of more resource-efficient foundational models.

Acknowledgments. The study was supported by the National Natural Science Foundation of China (81827805, 82130060, 61821002, 92148205), National Key Research and Development Program (2018YFA0704100, 2018YFA0704104), China Postdoctoral Science Foundation (2021M700772), Zhuhai Industry-University-Research Collaboration Program (ZH22017002210011PWC), Jiangsu Provincial Medical Innovation Center (CXZX202219), Collaborative Innovation Center of Radiation Medicine of Jiangsu Higher Education Institutions, Nanjing Life Health Science and Technology Project (202205045), and Jiangsu Province Science and Technology Support Project (BE2023769). The funding sources had no role in the writing of the report or the decision to submit the paper for publication.

Disclosure of Interests. The authors have no competing interests.

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
2. Gao, P., Lu, J., Li, H., Mottaghi, R., Kembhavi, A.: Container: Context aggregation network. arXiv preprint arXiv:2106.01401 (2021)
3. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
4. 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)
5. Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., Pedrycz, W.: A comprehensive survey on applications of transformers for deep learning tasks. Expert Systems with Applications p. 122666 (2023)
6. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022)
7. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International conference on machine learning. pp. 3519–3529. PMLR (2019)
8. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI

Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)

9. 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)
10. Liu, R., Li, Y., Tao, L., Liang, D., Zheng, H.T.: Are we ready for a new paradigm shift? a survey on visual deep mlp. Patterns **3**(7) (2022)
11. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
12. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
13. Lyu, P., Xiong, J., Fang, W., Zhang, W., Wang, C., Zhu, J.: Advancing multi-organ and pan-cancer segmentation in abdominal ct scans through scale-aware and self-attentive modulation. MICCAI FLARE2023 Challenge, (2023), https://openreview.net/forum?id=Mz7HMmc01M
14. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
15. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2021)
16. Nguyen, T., Raghu, M., Kornblith, S.: Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. arXiv preprint arXiv:2010.15327 (2020)
17. Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J.: Global filter networks for image classification. Advances in neural information processing systems **34**, 980–993 (2021)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
19. Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical Image Analysis p. 102802 (2023)
20. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)
21. Tatsunami, Y., Taki, M.: Sequencer: Deep lstm for image classification. Advances in Neural Information Processing Systems **35**, 38204–38217 (2022)
22. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. Advances in neural information processing systems **34**, 24261–24272 (2021)
23. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10819–10829 (2022)

24. Yu, X., Yang, Q., Zhou, Y., Cai, L.Y., Gao, R., Lee, H.H., Li, T., Bao, S., Xu, Z., Lasko, T.A., et al.: Unest: Local spatial representation learning with hierarchical transformer for efficient medical segmentation. Medical Image Analysis **90**, 102939 (2023)
25. Zhang, Z., Zhang, H., Zhao, L., Chen, T., Arik, S.Ö., Pfister, T.: Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3417–3425 (2022)