Edge-Aware Token Halting for Efficient and Accurate Medical Image Segmentation

Yuhao Guo[1,2], Bo Song[1], Heng Fan[3], and Erkang Cheng[1](✉)

[1] Institute of Intelligent Machines, HFIPS, Chinese Academy of Sciences, Hefei, China
ekcheng@iim.ac.cn
[2] University of Science and Technology of China, Hefei, China
[3] University of North Texas, Texas, USA

Abstract. The effectiveness of Vision Transformer (ViT)-based feature encoding network has been demonstrated in medical image analysis tasks. However, the complexity growing quadratically with the token number limits its application in dense prediction. To accelerate ViT, we propose an efficient and accurate token halting and reconstruction encoder framework, termed HRViT, designed for precise medical image semantic segmentation. Our approach is motivated by the observation that background and internal tokens can be easily identified and halted in early layers, while complex and ambiguous edge regions require deeper computational processing for accurate segmentation. HRViT leverages this insight by incorporating an edge-aware token halting module, which dynamically identifies edge patches and halts non-edge tokens. The preserved edge tokens are propagated to deeper layers and further refined through edge reinforcement. After encoding, all tokens are restored to their original positions, and auxiliary supervision is also introduced to strengthen the encoder's representation power. We evaluate the segmentation performance of our method using two public medical image datasets and the experimental results show that our method achieves promising performance compared with the state-of-the-art approaches. Our code is released at https://github.com/guoyh6/hrvit.

Keywords: Token Halting · Semantic Segmentation · Transformer.

1 Introduction

Recent works have shown impressive results using Transformer [23] for medical image segmentation [18,21]. Vision Transformers (ViTs) series [7,1,8,10] have gained recognition for their exceptional global dependency modeling capabilities, and have outperformed convolutional neural networks in various tasks. ViT divides an image into several patches and treats them as tokens. Due to the quadratic growth of time complexity in the Transformer block with token length, it becomes necessary to limit the number of patches based on large-scale 3D medical image processing tasks and device capabilities.

However, achieving accurate medical image segmentation necessitates the utilization of high-resolution information from the images, which often requires a substantial number of input tokens. Existing token reduction approaches mainly prune less informative tokens [19,24,25] or merge related tokens [4,13,14,15] at early Transformer layers, which is incompatible with dense prediction tasks. In DToP [22], the segmentation quality of all tokens is evaluated at each stage, and tokens with high confidence are pruned. However, in DToP, each token outputs independently, resulting in inferior segmentation quality. SCD [27] is a token sparsification model designed for medical image segmentation, where some tokens are halted during the encoding process. However, the halting method employed lacks explicit supervision, which can lead to suboptimal performance as the model may become confused about which tokens should be preserved.

To address the aforementioned issues, in this paper, we introduce HRViT, an efficient token **H**alting and **R**econstruction **ViT**-based method for medical imaging semantic segmentation. We propose a novel edge-aware halting module in encoder layers that guides the sparsity of tokens. This module is built on the observation of how background pixels and core regions of large objects naturally achieve higher confidence scores through early-layer segmentation heads [22]. The inherent ambiguity of semantic edges in medical images complicates precise segmentation, making it essential to preserve and propagate these regions. Tokens corresponding to the background or interior of objects can be easily recognized and halted in the early encoder layers. Identifying edge patches is simpler than predicting pixel-level semantic categories. Additionally, we introduce an edge reinforcement module to enhance the prediction reliability of selective critical edge tokens. After the encoding process, HRViT reassembles the full tokens by restoring both the edge tokens and the halted intermediate tokens. We also design an encoder auxiliary loss for restored tokens. Finally, all reconstructed tokens are utilized as input for the subsequent segmentation decoders [10,27] to compute the dense prediction results. Experiments show that HRViT achieves significantly faster inference speeds than baselines, with up to 46% and 78% improvement in FPS on two public benchmarks [12,2], while achieving state-of-the-art segmentation quality.

2 Method

2.1 Overview

The ViT-based 3D segmentation model follows a typical encoder-decoder architecture. The segmenter receives a cropping volume $I \in \mathbb{R}^{H\times W\times D\times C_{in}}$ initially divided into patches and flattened into a sequence of token embeddings $X \in \mathbb{R}^{N\times C}$, $N = \frac{H}{P} \times \frac{W}{P} \times \frac{D}{P}$ (P is patch size). The encoder processes the resultant tokens using L stacked Transformer block (TB) layers, while the decoder performs dense semantic prediction on the encoded embeddings. As shown in Fig. 1, our HRViT modifies the encoder component: an edge-aware token halting module, edge reinforcement, and a token reconstruction strategy.

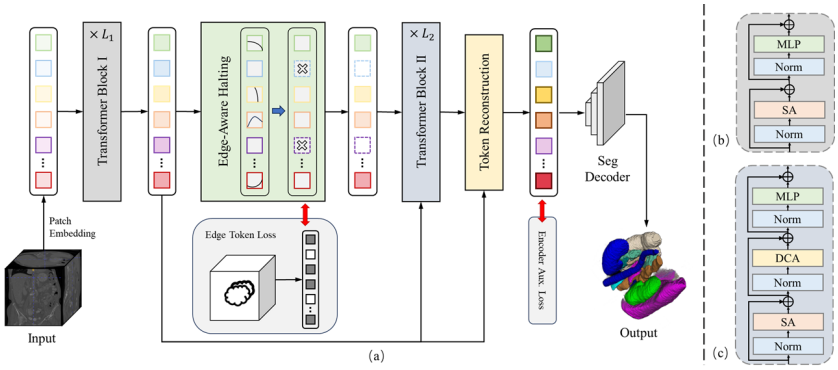

Fig. 1. Overview of the proposed HRViT for medical image segmentation. (b) Standard Transformer block (TB_I). (c) Transformer block with edge reinforcement (TB_{II}).

2.2 Edge-Aware Halting Module

HRViT partitions the L-layer encoder into two phases. In the first phase (L_1-layer), we retain the conventional strategy of processing all tokens $X \in \mathbb{R}^{N \times C}$. In the second phase ($L_2$-layer), we introduce an edge-aware token halting module that dynamically selects a small subset of tokens $X_{keep} \in \mathbb{R}^{M \times C}$ to propagate forward. This selection is based on an edge-aware halting score, where the number of tokens M is significantly smaller than the total number of tokens N, i.e., $M \ll N$. Therefore, it reduces the encoder's computational complexity while prioritizing semantically critical edge regions.

We employ a scoring network to determine which tokens should be halted. This scoring network predicts the edge confidence $S \in \mathbb{R}^{N \times 1}$ of all tokens, which is implemented using an MLP followed by a sigmoid activation. A pre-defined threshold t is then applied to determine which tokens to halt. The binary halting mask is obtained by applying a non-differentiable indicator function to edge confidence. We adopt the straight-through estimator (STE) [3] strategy to handle the non-differentiable thresholding function, treating it as an identity function during the gradient backpropagation. To encourage the scoring network to output desirable edge confidence S, we employ a cross-entropy (CE) loss function that aligns the predictions with edge supervision, i.e., $\mathcal{L}_{edge} = CE(S, GT_{token})$. The binary label of token $GT_{token} \in \mathbb{R}^N$ is labeled positive when its corresponding patch region exists semantic inconsistency, indicating edge tokens.

2.3 Edge Reinforcement

The semantic categories of edge tokens are strongly connected to the surrounding area. To enhance the prediction reliability of selective critical tokens, we retain halted tokens in subsequent encoder layers as semantic context providers. As illustrated in Fig. 1 (c), our approach differs from standard self-attention (Fig. 1 (b)) by incorporating a deformable cross-attention module [28] into each

Transformer block during the second encoding phase. Following the conventional self-attention operation, the selected tokens X_{keep} are utilized as queries, while the complete pre-halting token sequence X^{L_1} provides the keys and values. This design enables the selective tokens to adaptively sample and aggregate region features of interest, leveraging the semantic context to enhance representation learning. Mathematically, the encoding patterns can be described as follows:

$$\begin{cases} X^i = TB_I^i(X^{i-1}), & i \leq L_1 \\ X_{keep}^i = TB_{II}^i(X_{keep}^{i-1}, X^{L_1}), & i > L_1 \end{cases} \tag{1}$$

The additional computational cost is minimal, as the complexity of the deformable attention mechanism scales linearly with the number of queries M.

2.4 Token Reconstruction & Decoder

For the dense semantic segmentation task, we utilize a standard segmentation decoder that processes a full-length token sequence. To meet this requirement, we introduce a token reorganization module after the encoding phase. This module restores the sparse edge tokens and the halted intermediate tokens to their original positions in the initial sequence. Following the reconstruction, a Transformer block is applied to the restored encoding tokens to fully integrate the features from the two-stage encoding process, which can be written as: $X_{enc} = TB_I(Restore(X_{keep}^L, X_{halt}^{L_1}))$. The token reconstruction mechanism effectively mitigates performance degradation caused by the missing blur edge token. Our primary goal is to develop a ViT-based encoder acceleration framework that operates independently of the decoder workflow. Consequently, HRViT leverages existing segmentation decoders [10,26] to produce the final dense segmentation results.

2.5 Optimization

In addition to the standard segmentation loss, we introduce two patch-level auxiliary losses. The first is the edge loss, as described earlier, which is applied to the scoring network. The second is an auxiliary CE loss, denoted as $\mathcal{L}_{aux}$, applied to the restored encoding tokens. Each token is processed by an MLP head to predict whether it belongs to an edge region or a specific semantic category. The output dimension of the prediction is set to $cls + 1$, where cls represents the number of semantic classes, and the additional dimension corresponds to the edge category. The overall training loss is formulated as: $\mathcal{L} = \mathcal{L}_{edge} + \mathcal{L}_{aux} + \mathcal{L}_{seg}$, where $\mathcal{L}_{seg}$ consists of pixel-level CE loss and Dice loss. All loss weights are equal.

3 Experiment

3.1 Dataset

We evaluate HRViT on two publicly available 3D medical segmentation datasets.

BTCV Multi-Organ Segmentation dataset (CT). The BTCV (Multi Atlas Labeling Beyond The Cranial Vault) consists of 30 abdominal clinical CT scans [12]. Following the previously reported works [5,6], we use 18 volumes for training and the rest 12 volumes are used for testing. The average Dice Similarity Coefficient (DSC) and average Hausdorff Distance (HD) are reported to evaluate our method on 8 abdominal organs (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, stomach).

Brain Tumor Segmentation (MRI) is a subtask of The Medical Segmentation Decathlon [2]. The Brain Tumor Segmentation (BraTS) dataset provides 484 multi-modal multi-site (FLAIR, T1w, T1gd, T2w) MRI annotated cases. Following SCD [27], we split the dataset into training, validation, and testing sets in an 80%:15%:5% ratio.

3.2 Implementation details

In the experiments, we use a single NVIDIA A800 GPU with PyTorch and MONAI frameworks. We adapt SCD [27] for data processing and optimization strategy. For baseline comparisons, we select two representative ViT-based segmentation models: UNETR and SegViT. The patch size is set to $P = 8$ and the embedding channel is $C = 768$. The encoder consists of $L = 12$ Transformer blocks, with the proposed edge-aware token halting module inserted after the third block, i.e., $L_1 = 3$ and $L_2 = 9$. To further evaluate performance, we implement a fixed-quantity token retention variant, referred to as HRViT-S, where only the top-ρ tokens with the highest edge confidence are retained. Through ablation studies, we determine that a scoring threshold to $t = 0.75$ and keeping ratio $\rho = 0.1$ yield optimal performance for the dynamic and static operational modes, respectively. During training, we randomly crop a fixed-size image patch as the input. For inference, we employ a sliding window approach with a half-window overlap to process the entire volume. The input size is set to $96 \times 96 \times 96$ for the BTCV dataset and $128 \times 128 \times 128$ for the BraTS dataset.

3.3 Main Results

In Table 1, we present the evaluation results on the BTCV dataset. Our method not only accelerates ViT-based segmentation models but also achieves state-of-the-art performance in both average Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) metrics. Compared to the baseline UNETR [10], HRViT demonstrates a 2.79% improvement in DSC and a reduction of 5.69 mm in HD. HRViT-S achieves a 2.47% increase in average DSC and a decrease of 2.60 mm in HD. Applying the HRViT encoder variants to SegViT [26], a general-purpose semantic segmentation framework, also yields consistent performance improvements. Our approach shows more significant advancements in the HD evaluation metric, indicating that our edge-aware token halting and edge reinforcement modules contribute to improved edge predictions.

In Table 2, we evaluate the efficiency of our HRViT on the BTCV dataset. We measure the efficiency of our network by profiling the encoder throughput,

Table 1. Comparison with other methods on BTCV. The best and second-best results are colored **red** and **blue**. Gallbladder: Gb; KL: Kidney (L); KR: Kidney (R).

Methods	DSC↑	HD↓	Aorta	Gb	KL	KR	Liver	Pancreas	Spleen	Stomach
V-Net [17]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR [9]	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
TransUNet [6]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUNet [5]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
SCD [27]	82.18	19.85	**89.39**	**73.60**	**85.66**	83.65	95.59	62.17	88.84	77.37
MetaUNETR [16]	82.10	-	-	-	-	-	-	-	-	-
nnU-Net Revisited [11]	**85.04**	-	-	-	-	-	-	-	-	-
SegViT [26]	79.63	10.99	82.69	66.67	82.02	79.57	96.06	58.06	**91.75**	78.21
HRViT-S+SegViT	80.47	**9.82**	84.21	66.90	81.33	80.09	95.45	63.37	90.85	79.54
HRViT+SegViT	80.35	**9.20**	83.05	68.21	82.12	78.37	**95.98**	63.44	90.61	**79.63**
UNETR [10]	80.74	17.38	88.93	67.71	84.30	**83.81**	**95.74**	58.82	89.36	75.46
HRViT-S+UNETR	83.21	14.78	**89.39**	70.84	85.58	**84.83**	95.36	**66.29**	**92.53**	79.32
HRViT+UNETR	**83.53**	11.69	89.11	**73.84**	**85.75**	83.48	95.59	**68.75**	89.87	**80.54**

Table 2. Segmentation efficiency of different methods on BTCV.

Methods	DSC↑	HD↓	Encoder Throughput(img/s)	FPS(img/s)	FLOPs(GMac)
UNETR [10]	80.74	17.38	51.26	34.68	273.45
SCD [27]	82.18	19.85	105.32	47.43	146.63
HRViT-S+UNETR	83.21	14.78	104.46	47.63	159.51
HRViT+UNETR	83.53	11.69	120.23	50.65	143.69

frames per second (FPS), and floating-point operations (FLOPs). HRViT is a dynamic network that allows for adjusting token numbers and computation based on the inputs. We conduct tests on multiple inputs and calculate the average as the final metric. Our approach achieves higher accuracy while significantly improving inference efficiency on the BTCV dataset. Specifically, HRViT achieves a 2.35× increase in encoder throughput and accelerates FPS from 34.68 images per second (imgs/s) to 50.65 imgs/s, representing a 46% improvement compared to the baseline.

Table 3. Segmentation accuracy and efficiency of different methods on BraTS.

Methods	DSC↑	HD↓	Encoder Throughput(img/s)	FPS(img/s)	FLOPs(GMac)
U-Net [20]	75.84	9.17	-	-	-
TransUNet [6]	73.24	11.74	-	-	-
UNETR [10]	77.68	7.96	15.69	11.66	824.38
SCD [27]	75.79	8.31	**47.54**	**20.44**	**428.28**
HRViT-S+UNETR	**78.68**	**7.59**	45.04	19.95	442.82
HRViT+UNETR	**78.39**	**7.90**	**48.12**	**20.74**	**407.78**

Table 3 provides qualitative brain tumor segmentation comparisons. UNETR [10] outperforms both CNN and transformer-based approaches on the BraTS dataset. Our HRViT-S further improves the average DSC by 1% and reduces the HD by 0.37 mm when using UNETR with a patch size of 8 as the baseline. In terms of efficiency, HRViT achieves an encoder throughput of

3.07× and accelerates the FPS from 11.66 images per second (imgs/s) to 20.74 imgs/s, resulting in a significant 78% improvement compared to the baseline. Our method also yields superiority over SCD [27] in both efficiency and accuracy.

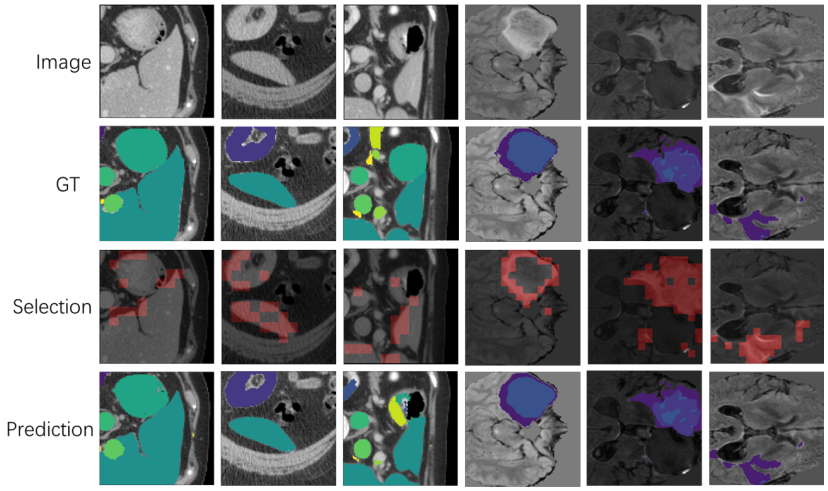

Fig. 2. Visualization results of HRViT on BTCV (left three columns) and BraTS (right three columns).

We provide visualizations of the halting policy and segmentation predictions for both datasets in Figure 2. In the third row, the patches marked in red represent the edge tokens retained by HRViT. These selected patches are primarily focused on the contours of the segmentation objects, allowing the model to improve its performance with only a small number of tokens. Our segmentation predictions along the contours demonstrate the ability of HRViT to capture fine edges with precision and accuracy. This ability is further validated by the evaluation metric of Hausdorff Distance (HD).

3.4 Ablation Studies

We conduct ablation studies on the UNETR baseline using the BTCV Multi-Organ Segmentation dataset.

Effects of the Proposed Modules. To investigate the effects of the edge-aware token halting module, edge reinforcement module, and encoder auxiliary loss, several experiments are conducted and the results are listed in Table 4. Compared to randomly halting 90% tokens, our edge-aware token selection criterion improves DSC by 2.07% and decreases HD by 5.82 mm (Row 1 and Row 3). Compared to a learnable token halting approach without additional supervision, our edge-aware token selection improves DSC by 1.02% and decreases HD by 3.61 mm (Row 2 and Row 3). We explore the impact of the edge score

Table 4. Ablation on the proposed modules. Acc_{edge} represents the overall prediction accuracy of whether tokens belong to edge regions.

	Halting Method	ρ / t	Reinforcement	$\mathcal{L}_{aux}$	Acc_{edge} ↑	DSC↑	HD↓
1	Random	$\rho = 0.1$			-	80.47	18.66
2	Non-Supervision	$\rho = 0.1$			-	81.50	16.45
3	Edge Supervision	$\rho = 0.1$			-	82.52	12.84
4	Edge Supervision	t=0.50			0.935	82.21	14.33
5	Edge Supervision	t=0.75			0.953	82.65	13.47
6	Edge Supervision	t=0.90			0.941	82.43	14.73
7	Edge Supervision	t=0.75		✓	0.962	82.95	13.07
8	Edge Supervision	t=0.75	✓		0.943	83.18	15.08
9	Edge Supervision	t=0.75	✓	✓	0.958	83.53	11.69

threshold t in HRViT and observe that the best value is 0.75 compared to 0.5 and 0.9 (Rows $4-6$). The experimental results demonstrate that a low threshold in the edge score prediction network results in imprecise edge detection, whereas a high threshold excessively restricts the number of tokens identified as edge tokens. The proposed edge reinforcement module improves the performance of DSC from 82.65% to 83.18% in HRViT+UNETR. We obtain similar observations on encoder auxiliary loss, the additional loss brings a 0.3% improvement on DCS.

Table 5. Ablation on the keeping ratio ρ on HRViT-S. The experiments only compare different variants of the halting module without edge reinforcement and auxiliary loss.

ρ	DSC↑	HD↓	Encoder Throughput(img/s)	FPS(img/s)	FLOPs(GMac)
1.00	80.74	17.38	51.26	34.68	273.45
0.50	82.27	15.37	85.78	43.38	206.09
0.25	82.42	14.56	113.58	49.49	170.75
0.10	82.52	12.84	125.78	51.88	151.99

Effects of the Keeping Ratio on HRViT-S.

The performance and inference speed with different ρ are listed in Table 5. Our results show that performance and efficiency improve as the keeping ratio decreases to 0.1, indicating that focusing computation on the top 10% of highest-scoring tokens enhances both accuracy and speed.

4 Conclusion

In this paper, we propose HRViT, a ViT-based edge-aware token halting method for 3D medical image segmentation. HRViT selects edge tokens and passes them to deeper Transformer blocks. The rest tokens are halted encoding and referenced as keys to enhance the selected tokens. Before dense decoding, we reconstruct the complete sequence and introduce the encoder auxiliary loss to improve segmentation performance. Our HRViT outperforms the baseline even when halting

majority encoder tokens, and achieves 1.46× and 1.78× FPS on two public medical semantic segmentation benchmarks respectively.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (62163011); Anhui Provincial Key Research and Development Program (2023s07020017); and Anhui Provincial Key Laboratory of Bionic Sensing and Advanced Robot Technology. Heng Fan is not supported by any funds for this work.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Almalik, F., Yaqub, M., Nandakumar, K.: Self-ensembling vision transformer (sevit) for robust medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 376–386. Springer (2022)
2. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
3. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)
4. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your ViT but faster. In: International Conference on Learning Representations (2023)
5. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
6. 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)
7. 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)
8. Du, S., Bayasi, N., Hamarneh, G., Garbi, R.: Mdvit: Multi-domain vision transformer for small medical image segmentation datasets. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 448–458. Springer (2023)
9. Fu, S., Lu, Y., Wang, Y., Zhou, Y., Shen, W., Fishman, E., Yuille, A.: Domain adaptive relational reasoning for 3d multi-organ segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 656–666. Springer (2020)
10. 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)

11. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 488–498. Springer (2024)
12. 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)
13. Liang, W., Yuan, Y., Ding, H., Luo, X., Lin, W., Jia, D., Zhang, Z., Zhang, C., Hu, H.: Expediting large-scale vision transformer for dense prediction without fine-tuning. Advances in Neural Information Processing Systems **35**, 35462–35477 (2022)
14. Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not all patches are what you need: Expediting vision transformers via token reorganizations. In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=BjyvwnXXVn_`
15. Lu, C., de Geus, D., Dubbelman, G.: Content-aware token sharing for efficient semantic segmentation with vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23631–23640 (2023)
16. Lyu, P., Zhang, J., Zhang, L., Liu, W., Wang, C., Zhu, J.: Metaunetr: Rethinking token mixer encoding for efficient multi-organ segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 446–455. Springer (2024)
17. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
18. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: A robust volumetric transformer for accurate 3d tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 162–172. Springer (2022)
19. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems **34**, 13937–13949 (2021)
20. 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)
21. She, D., Zhang, Y., Zhang, Z., Li, H., Yan, Z., Sun, X.: Eoformer: Edge-oriented transformer for brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 333–343. Springer (2023)
22. Tang, Q., Zhang, B., Liu, J., Liu, F., Liu, Y.: Dynamic token pruning in plain vision transformers for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 777–786 (2023)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
24. Ye, M., Meyer, G.P., Chai, Y., Liu, Q.: Efficient transformer-based 3d object detection with dynamic token halting. arXiv preprint arXiv:2303.05078 (2023)

25. Yin, H., Vahdat, A., Alvarez, J.M., Mallya, A., Kautz, J., Molchanov, P.: A-vit: Adaptive tokens for efficient vision transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10809–10818 (2022)
26. Zhang, B., Tian, Z., Tang, Q., Chu, X., Wei, X., Shen, C., et al.: Segvit: Semantic segmentation with plain vision transformers. Advances in Neural Information Processing Systems **35**, 4971–4982 (2022)
27. Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P.: Token sparsification for faster medical image segmentation. In: International Conference on Information Processing in Medical Imaging. pp. 743–754. Springer (2023)
28. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)