Revisiting Self-Attention in Medical Transformers via Dependency Sparsification

Xian Lin[1], Zhehao Wang[1], Zengqiang Yan[1(✉)], and Li Yu[1]

School of Electronic Information and Communications, Huazhong University of Science and Technology
{xianlin, zhehao_wang, z_yan, hustlyu}@hust.edu.cn

Abstract. Vision transformer (ViT), powered by token-to-token self-attention, has demonstrated superior performance across various vision tasks. The large and even global receptive field obtained via dense self-attention, allows it to build stronger representations than CNN. However, compared to natural images, both the amount and the signal-to-noise ratio of medical images are small, often resulting in poor convergence of vanilla self-attention and further introducing non-negligible noise from extensive unrelated tokens. Besides, token-to-token self-attention requires heavy memory and computation consumption, hindering its deployment onto various computing platforms. In this paper, we propose a dynamic self-attention sparsification method for medical transformers by merging similar feature tokens for dependency distillation under the guidance of feature prototypes. Specifically, we first generate feature prototypes with genetic relationships by simulating the process of cell division, where the number of prototypes is much smaller than that of feature tokens. Then, in each self-attention layer, key and value tokens are grouped based on their distance from feature prototypes. Tokens in the same group, together with the corresponding feature prototype, would be merged into a new prototype according to both feature importance and grouping confidence. Finally, query tokens build pair-wise dependency with such newly-updated prototypes for fewer but global and more efficient interactions. Extensive experiments on three publicly available datasets demonstrate the effectiveness of our solution, working as a plug-and-play module for joint complexity reduction and performance improvement of various medical transformers. Code is available at https://github.com/xianlin7/DMA.

Keywords: Medical transformer · Efficient self-attention · Sparse dependency · Feature prototypes.

1 Introduction

Vision transformer (ViT) has exhibited exceptional performance across various computer vision tasks and attracted widespread concern in medical image analysis [1]. In contrast to convolution neural networks (CNN) focusing on translation invariance and locality, ViT adopts a distinctive paradigm by gridding an input

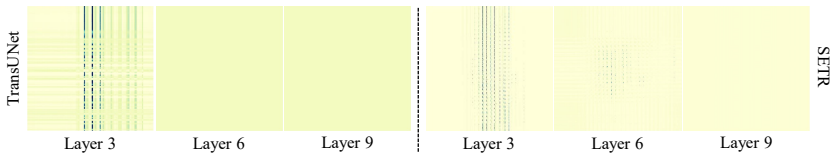


Fig. 1: Vanilla self-attention matrices trained on BTCV where attention maps tend to be uniform in deeper layers.

image into a series of tokens and building pair-wise global interactions through self-attention [2]. Specifically, self-attention updates each token by aggregating all tokens with dependency weights [3]. In this way, self-attention brings a global receptive field to ViT, promoting its superior feature representation ability [4].

Every coin has two sides. The calculation of self-attention is quadratic to the number of tokens, resulting in daunting computational complexity and memory consumption and in turn hindering the deployment of ViT on clinical equipment [5, 6]. Inspired by attention sparsification in natural scenes [7, 8], several backbones designed for medical imaging have adopted similar strategies to simplify self-attention. As two representative approaches, Swin-Unet [9] adopts window-based local attention while MISSFormer down-samples the key and value feature tokens by fusing tokens in the same grid window into a new token [10]. Though such hand-crafted approaches efficiently reduce computational complexity and memory consumption [11], the lack of medical context awareness may result in the loss of important clinical features during dependency reduction and thereafter performance degradation.

To analyze the performance of ViT in medical imaging, we trained ViT-based models on a commonly-used abdominal multi-organ dataset (*i.e.*, BTCV [12]) and visualized the learned attention matrices as presented in Fig. 1. Like typical medical datasets, the number of slices/images in BTCV is fewer than $5k$ and the average foreground proportion across different organs is lower than 6% (*as stated in the supplementary material*). In other words, compared to natural scenes, medical datasets are of small scale and low signal-to-noise ratios. As depicted in Fig. 1, when training ViT on such datasets, as the network goes deep, attention matrices gradually tend to become uniform, losing the ability to recognize important dependencies. On the one hand, larger attention matrices require more data to build effective dependencies. On the other hand, larger background proportions bring more interference from irrelevant tokens/regions, resulting in sub-optimal global dependencies. Therefore, simplifying self-attention calculation while increasing the signal-to-noise ratio in dependency is crucial for unleashing the potential of ViT in medical imaging.

In this paper, we revisit self-attention in ViTs from the perspective of dependency sparsification, aiming at increasing the signal-to-noise ratio of medical imaging for performance improvement while reducing both computational complexity and memory consumption. Specifically, we propose a plug-and-play dependency merging attention (DMA) mechanism to conveniently boost vanilla

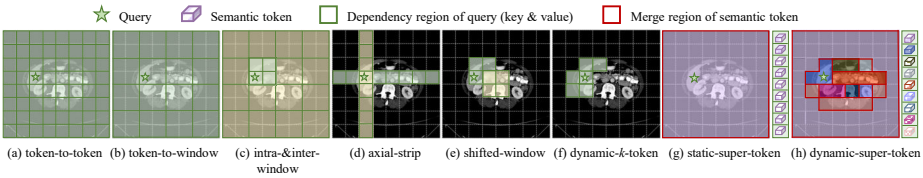


Fig. 2: Vanilla self-attention and its sparse variants.

ViT for better attention convergence and lower deployment cost. Based on feature prototypes, generated by imitating the process of cell division, DMA dynamically divides feature tokens into groups centered around different feature prototypes. By merging tokens within the same group into a new prototype based on both feature importance and grouping confidence, such newly-formed prototypes can well replace original feature tokens to provide global information for queries. As the number of prototypes is much smaller than that of feature tokens, as depicted in Fig. 2 (h), a long series of feature tokens are adaptively merged into several regions in DMA. In this way, dependency is more sparse and the foreground proportion is further increased, providing stronger keys and values for attention calculation. Comprehensive experiments on various tasks and backbones demonstrate the effectiveness of DMA as a plug-and-play module for joint complexity reduction and performance improvement.

2 Related Works

Attention Sparsification in ViTs. As illustrated in Fig. 2, existing research on attention sparsification can be categorized into token-to-window, intra-&-inter-window, axial-strip, shifted-window, dynamic-k-token, static-super-token, and dynamic-super-token. Token-to-window methods down-sample keys and values via a similar operation with uniform tokenization [13, 14]. Intra-&-inter-window methods build dependency within and across windows sequentially [15]. Axial-strip methods achieve information transmission by alternately establishing dependencies on different strips [16]. Shifted-window methods build interactions within a local window [17]. Dynamic-k-token approaches only focus on k important regions by predicting the biases or importance scores [8, 7, 18]. By learning a transformation matrix, methods based on static super tokens fuse all feature tokens with varying weight assignments into different super tokens [19, 20]. The proposed DMA can be viewed as a dynamic-super-token approach, which fuses different regions into different super tokens based on their semantic similarity. Compared to existing mechanisms, DMA focuses on adaptively merging redundant dependencies and increasing the proportion of interested regions.

ViTs in Medical Imaging. Encouraged by the great success of ViT in computer vision [1], transformer-based models have sparked a research boom in medical image analysis [5]. TransUNet is the first transformer-based model proposed for medical image segmentation [31]. Inspired by this, a series of frameworks

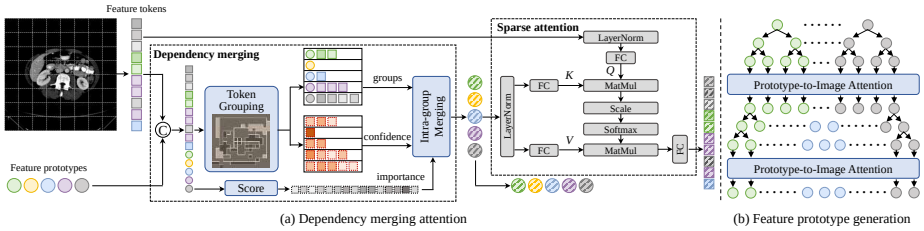


Fig. 3: Illustration of dependency merging attention.

have been proposed to combine ViT with CNN for stronger feature representation learning [21, 22]. However, the heavy computational complexity and memory consumption limit the application of ViT in medical scenarios. Window-based and axial-based transformers (as shown in Fig. 2 (e) and (d)) are applied to balance deployment costs and receptive fields [9, 23]. To pursue device-friendly global receptive fields, MISSFormer conducts dependency reduction by fusing the keys and values located in the same grid [10]. Such sparsification methods are either hand-crafted or locality-based, instead of content-aware global sparsification. Comparatively, the proposed DMA can realize global perception by dynamically merging similar tokens for fewer dependencies and a higher signal-to-noise ratio.

3 Method

The Overall Architecture. DMA is illustrated in Fig. 3, which adaptively merges dependency carrying similar information under the assistance of feature prototypes to realize sparse attention. Specifically, we first condense the stable structure, texture, and distribution of objects into initial feature prototypes. In each self-attention layer, taking each prototype as a cluster center, feature tokens (*i.e.*, vanilla tokens in ViT) are divided into different groups according to their feature similarity with each feature prototype. Then, tokens from the same group are fused into a new prototype weighted by both grouping confidence and feature importance. By regarding the newly-fused prototypes as key-value pairs, each query can build global dependency with fewer costs and less noise interference. The above process contains three key steps, including feature prototype generation, token grouping, and intra-group merging, each of which is described in detail in the following. *The deployment of DMA in ViTs is illustrated in the supplementary material.*

Feature prototype generation. Feature prototypes are generated from k initial meta prototypes $P_0 \in \mathbb{R}^{k\times d}$ (*i.e.*, the 0-th level) by imitating the process of cell division as depicted in right of Fig. 3. Meta prototypes are input-independent feature embeddings, which are learned from all training data and represent the stable and common characteristics of organs/lesions. Each meta prototype $P_0^i \in \mathbb{R}^{1\times d}, i \in k$ would generate two new prototypes P_1^{2i} and P_1^{2i+1}

in division, formulated by

$$[P_1^{2i}, P_1^{2i+1}] = P_0^i E_1^i, \tag{1}$$

where $E_1^i \in \mathbb{R}^{d\times 2d}$ is the projection matrix for the i-th prototype in the first division and $P_1^{2i} \in \mathbb{R}^{1\times d}$ and $P_1^{2i+1} \in \mathbb{R}^{1\times d}$ are the newly-generated prototypes. After repeating the above generation process for four times, a series of feature prototypes $P_4 \in \mathbb{R}^{M\times d}$ are generated, where $M = 16k$. To adapt such common/shared features (*i.e.*, feature prototypes) to each input image, we introduce prototype-to-image attention every two divisions, formulated as

$$A_l = \sigma(\mathbf{LN}(P_l)W_q(\mathbf{LN}(F)W_k)^T/\sqrt{d}), \tag{2}$$

$$F_a = (A_l\mathbf{LN}(F)W_v)W_l + P_l, \tag{3}$$

$$P_l' = \mathbf{MLP}(\mathbf{LN}(F_a)) + F_a, \tag{4}$$

where $F \in \mathbb{R}^{N\times d}$, P_l' $(l = 2, 4)$, and σ represent feature tokens, the updated prototypes, and Softmax, W_q, W_k, W_v, and W_l are projection matrices, and **LN** is Layer Normalization. Finally, the initial feature prototypes are generated as $P = P_4'$. By generating feature prototypes through cell division, both shallow feature similarity and deep semantic heterogeneity of objects are well preserved.

Token grouping. Given feature prototypes $P \in \mathbb{R}^{M\times d}$, tokens in $F \in \mathbb{R}^{N\times d}$ are grouped by calculating their Euclidean distances with P. Specifically, for better computation efficiency, feature prototypes P and tokens F are concatenated as $\mathcal{F} = [P, F] \in \mathbb{R}^{(M+N)\times d}$ and the distance between the i-th row/element $\mathcal{F}(i) \in \mathbb{R}^{1\times d}$ and the j-th feature prototype $P(j) \in \mathbb{R}^{1\times d}$ is defined as $D(i, j) = ||\mathcal{F}(i) - P(j))||_2/\sqrt{d}$, where $D \in \mathbb{R}^{(M+N)\times M}$ reflects the feature similarity between each row/element in $\mathcal{F}$ and feature prototypes. Taking the index $I \in \mathbb{R}^{(M+N)}$ (*i.e.*, $I(i) \in [1, M]$) of the smallest value in each row of D as its grouping category, elements with similar features are grouped. As the closest prototype of each prototype is itself (*i.e.*, $D(i, i) = 0, i \in [1, M]$), each prototype is assigned to its original group and regarded as the cluster/group center. In this way, each group would have at least one element.

Intra-group merging. After token grouping, for any prototype $P(j)$, elements from the same group, *i.e.*, $\mathcal{F}(\phi(I, j)) \in \mathbb{R}^{G\times d}$ where $\phi(I, j)$ represents the indexes of I with the value of j and G is the number of elements in the j-th group, are merged to update $P(j)$ through a weighted combination based both grouping confidence $C_j \in \mathbb{R}^{1\times G}$ and feature importance $S_j \in \mathbb{R}^{1\times G}$. C_j reflects the degree of certainty during grouping, *i.e.*, the higher the grouping confidence of an element when its distance to the j-th group is much smaller than those to other groups, calculated by $C_j = C(\phi(I, j), j)/\sum_{m=1}^{M} C(\phi(I, j), m)$, where $C(i, j) = e^{-D(i,j)}/\sum_{m=1}^{M} e^{-D(i,m)}$. By emphasizing more on those elements with higher grouping confidence, updated prototypes would be biased towards the most representative features. S_j reflects the importance of a feature embedding, which is calculated by $S_j = S(\phi(I, j))/\sum S(\phi(I, j))$, where $S = 1/(1 + e^{-\mathcal{F}W_s})$ with projection $W_s \in \mathbb{R}^{d\times 1}$. By up-weighting the elements

with higher importance, crucial information would be enhanced. Finally, $P(j)$ is updated by $P(j) \leftarrow \sum(C_j\mathcal{F}(\phi(I,j)) + S_j\mathcal{F}(\phi(I,j)))/2$.

Prototype Loss. To obtain high-quality prototypes, we further divide prototypes into negative $P_- = \{P(j)|j < \frac{M}{2}\}$ and positive $P_+ = \{P(j)|j \geq \frac{M}{2}\}$, with P_+ representing the common features of target objects or key anatomical structures and P_- representing the common features of background or objects similar to the background. On the one hand, the feature distance between P_+ and P_- is expected to be larger than any distance between prototypes within P_+ or P_-, making foreground and background features more distinguishable. On the other hand, the feature distance within P_+ or P_- is expected to be small to ensure the accuracy of token merging but should not be too small to ensure prototype diversity. Therefore, we construct a prototype loss $\mathcal{L}_p$ defined as

$$\mathcal{L}_{p_+} = \|\bar{P_+} - \bar{P}\|_2 - \mathbf{Max}(\mathbf{Avg}(\|P_+ - \bar{P_+}\|_2), \beta), \tag{5}$$

$$\mathcal{L}_{p_-} = \|\bar{P_-} - \bar{P}\|_2 - \mathbf{Max}(\mathbf{Avg}(\|P_- - \bar{P_-}\|_2), \beta), \tag{6}$$

$$\mathcal{L}_p = \mathbf{Max}(\alpha - 0.5L_{p_+} - 0.5L_{p_-}, 0), \tag{7}$$

where β and α are trade-off hyper-parameters. $\bar{P}$, $\bar{P_+}$, and $\bar{P_-}$ are the average of P, P_+, and P_-, respectively. The smaller the β, the higher the similarity within P_+ and P_-. The larger the α, the farther the distance between P_+ and P_-.

Complexity Analysis of DMA. The computation of DMA consists of dependency merging $\Omega(\text{DM})$ and sparse self-attention $\Omega(\text{SSA})$:

$$\Omega(\text{DM}) = (3M + 4d)N + dM^2 + 3dM, \tag{8}$$

$$\Omega(\text{SSA}) = (2M + 2d)chN + 2chdM. \tag{9}$$

As $M \ll N$, the complexity of DMA is $O(N)$, which is smaller than that of vanilla self-attention (*i.e.*, $O(N^2)$). The computation of prototype generation is:

$$\Omega(\text{PG}) = (\frac{5}{2}M + 6d)hdN + \frac{5}{4}hd^2M + 10d^2M. \tag{10}$$

As all DMA layers share the same prototype generation in a backbone, the computation in prototype generation can be ignored compared to self-attention. Finally, the complexity of self-attention can be reduced from $\Omega(N^2)$ to $\Omega(N)$.

4 Experiments

Datasets. Three publicly available datasets are selected for evaluation. **(1) BTCV**[1]**.** An abdominal CT dataset consists of 30 scans with 13 annotated organs [12]. **(2) INSTANCE**[2]**.** A dataset contains 100 publicly available brain CT scans with pixel-wise annotations of intracranial hemorrhage [25]. **(3) ACDC**[3]**.**

[1] https://www.synapse.org/#!Synapse:syn3193805/wiki/217789/
[2] https://instance.grand-challenge.org/
[3] https://www.creatis.insa-lyon.fr/Challenge/acdc/

Table 1: Quantitative results (*i.e.*, Dice (%)) by replacing self-attention in TransUNet with different efficient-attention methods on BTCV. AG denotes the combination of right adrenal gland (LAG) and left adrenal gland (RAG).

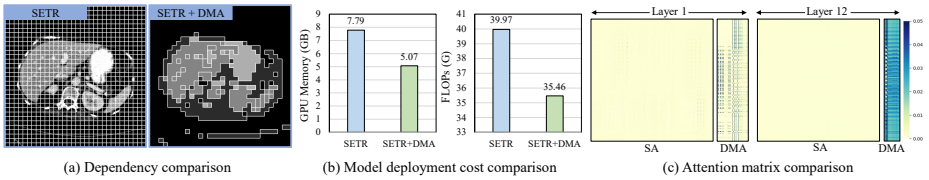
Method	Spl	Rkid	Lkid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	Avg
TransUNet [31]	91.35	83.67	86.83	58.89	69.49	95.99	75.29	90.34	83.86	68.55	68.67	57.91	76.06
+ DA [8]	90.28	84.07	87.43	58.27	69.42	95.65	76.20	90.34	83.85	69.67	68.46	61.69	76.69
+ Axial [16]	91.34	85.23	89.74	59.79	70.61	95.95	77.57	89.91	83.93	69.18	69.57	58.69	76.94
+ SRA [13]	87.93	84.61	89.99	56.80	71.81	95.86	75.48	90.10	83.25	69.82	70.11	62.31	76.95
+ kNN [7]	89.68	85.94	89.34	60.74	71.25	95.96	78.76	90.44	83.32	67.41	67.16	60.83	77.05
+ SSSA [15]	90.73	85.29	87.56	62.16	68.85	95.67	78.58	90.59	84.69	67.96	68.32	61.71	77.21
+ Swin [17]	91.27	85.10	**90.05**	56.36	68.89	95.91	79.94	89.71	84.25	69.32	70.21	61.26	77.19
+ PaCa [20]	90.58	84.74	89.94	62.00	71.79	95.78	77.49	89.69	83.43	69.71	67.73	60.25	77.18
+ SSA [14]	90.79	84.11	89.51	58.16	70.73	95.77	77.91	90.72	83.24	70.10	70.56	**62.77**	77.47
+ STA [19]	91.35	85.43	89.46	55.60	71.04	95.93	78.89	90.08	83.54	71.42	71.13	61.12	77.39
+ BRA [18]	91.76	85.57	87.99	60.62	71.16	95.94	80.17	90.97	84.66	68.09	71.07	59.63	77.48
+ **DMA**	**92.23**	**86.62**	89.00	**63.10**	**72.30**	**96.03**	**81.68**	**91.08**	**85.12**	**71.60**	**71.28**	61.31	**78.67**

An automated cardiac diagnosis dataset consists of 100 scans [24]. Following the setting in [31, 26], the partitioning ratios of 18/12 and 70/10/20 are utilized to split BTCV and ACDC. INSTANCE is randomly divided into the training, validation, and testing sets according to a ratio of 7:1:2.

Implementation Details. The models were implemented in PyTorch 1.8.0. under the same settings, *i.e.*, an Adam optimizer with an initial learning rate of 0.0001 and a batch size of 8 for 400 rounds.

Comparison with Efficient-Attention Methods. Totally 10 state-of-the-art (SOTA) efficient-attention methods are used to replace vanilla self-attention in TransUNet [31] for comparison as summarized in Table 1. In general, reducing dependency redundancy is beneficial for performance improvement. Among comparison methods, BRA achieves the best performance with an average increase of 1.42% in Dice, while Swin and SSA outperform all other methods in the segmentation of Lkid and AG, respectively. Comparatively, the proposed DMA achieves the best segmentation performance on 10 out of 13 organs, leading to the best overall segmentation performance and outperforming BRA and baseline with an average increase of 1.19% and 2.61% respectively in Dice.

Comparison with Segmentation Methods. We insert DMA into five ViT-based methods and conduct quantitative comparisons with 13 SOTA segmentation methods across three datasets. As stated in Table 2, both 2D (*i.e.*, SETR, TransFuse, TransUNet, and FAT-Net) and 3D (*i.e.*, TransBTS) backbones with vanilla transformer layers benefit from DMA for performance improvement. Furthermore, the performance of TransBTS+DMA and FAT-Net+DMA surpasses all comparison methods on BTCV&INSTANCE and ACDC, respectively.



(a) Dependency comparison (b) Model deployment cost comparison (c) Attention matrix comparison

Fig. 4: Comparison results between SETR [4] and SETR+DMA.

Table 2: Quantitative results (*i.e.*, Dice (%)) of various methods equipped with DMA against SOTA methods.

Method	BTCV	ACDC	INSTANCE
U-Net [27]	79.78	89.41	73.26
CA-Net [28]	77.80	91.95	73.74
AA-Unet [29]	78.70	91.46	73.21
MISSFormer [10]	75.12	91.19	72.62
H2Former [32]	77.57	92.40	73.72
SETR [4]	64.09	87.14	67.96
SETR+DMA	67.62	88.58	69.57
TransFuse [21]	60.28	89.10	64.42
TransFuse+DMA	61.43	90.39	67.89
TransUNet [31]	76.06	90.80	68.20
TransUNet+DMA	78.66	92.01	70.25
FAT-Net [22]	78.35	91.46	72.20
FAT-Net+DMA	80.44	**92.67**	73.39
nnU-Net [30]	83.36	91.61	71.12
nnFormer [26]	81.15	92.06	71.47
MedNeXt [33]	82.11	89.00	70.67
TransBTS [34]	82.42	90.59	74.06
TransBTS+DMA	**84.12**	91.23	**77.11**

Table 3: Ablation study on the number of prototypes M and the down-sampling frequency f on TransUNet evaluated on BTCV.

M	f	Dice	HD	IoU	SE	GFLOPs	Param.
32	6	77.57	20.00	66.63	78.56	**29.89**	112.07
64	**6**	**78.67**	19.25	**67.93**	**79.67**	30.50	112.07
128	6	77.65	19.29	66.78	78.32	31.72	112.07
64	3	77.92	19.47	66.98	77.37	30.24	112.07
64	4	78.22	19.81	67.41	78.32	30.35	112.07
64	12	77.37	**19.21**	66.6	76.63	30.73	112.07

Table 4: Component-wise ablation study of prototype generation. P-t-I is short for prototype-to-image attention.

division	P-t-I	L_p	Dice	HD	IoU	SE
○	○	○	76.87	20.35	65.94	78.03
●	○	○	77.58	19.67	66.51	78.30
●	●	○	77.92	**18.99**	67.14	77.68
●	●	●	**78.67**	19.25	**67.93**	**79.67**

Ablation Study. Comparison results under various settings of M and f are summarized in Table 3. Given a smaller M, more similar tokens would be merged, resulting in lower computational complexity but possible information loss, especially for small-size objects. Comparatively, given a larger M, some prototypes may be redundant and work as a global messenger, bringing negative effects from irrelevant tokens. f is the down-sampling frequency of prototypes (described in supplementary materials). Merging adjacent prototypes (*i.e.*, down-sampling) is to enrich the semantic features of prototypes and lower the computational complexity. But it may also be harmful for small-size objects as discussed above. Therefore, the selection of M and f is task-specific. Component-wise ablation studies in Table 4 validate the designs of prototype generation.

Visualization. Qualitative comparison before and after introducing DMA to SETR [4] on BTCV is illustrated in Fig. 4. Through DMA, similar tokens are

merged to reduce redundant dependency, thereby leading to lower GPU memory and computational costs and smaller but richer attention matrices.

5 Conclusion

We revisit self-attention in medical transformers and propose dependency merging attention (DMA) for joint complexity reduction and performance improvement. In DMA, similar feature tokens are adaptively merged under the guidance of feature prototypes, which greatly reduces token redundancy from either background or repetitive foreground regions. Experiments across various backbones on three datasets validate the effectiveness of DMA as a plug-and-play module.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62271220 and Grant 62202179, in part by the Natural Science Foundation of Hubei Province of China under Grant 2022CFB585, and in part by the Fundamental Research Funds for the Central Universities, HUST: 2024JYCXJJ032. The computation is supported by the HPC Platform of HUST.

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

References

1. Han, K., et al.: A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 87–110 (2022)
2. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Vaswani, A., et al.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
4. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6881–6890 (2021)
5. Li, J., et al.: Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. Med. Image Anal. **85**, 102672 (2023)
6. Shamshad, F., et al.: Transformers in medical imaging: A survey. Med. Image Anal. **88**, 102802 (2023)
7. Wang. P., et al.: Going deeper with image transformers. In: European Conference on Computer Vision, pp. 285–302 (2022)
8. Xia, Z., Pan, X., Song, S., Li, L. E., Huang, G.: Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4794–4803 (2022)
9. Cao, H., et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218 (2022)
10. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: MISSFormer: An effective transformer for 2d medical image segmentation. IEEE Trans. Med. Imag. **42**(5), 1484–1494 (2022)

11. Ou, Y., et al.: Patcher: Patch transformers with mixture of experts for precise medical image segmentation. In: Wang, Li., Dou, Q., Fletcher, P.T., Speidel S., Li, S. (eds.) MICCAI 2022, LNCS, vol. 13431, pp. 475–484. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_46
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, pp. 12 (2015)
13. Ren, S., Zhou, D., He, S., Feng, J., Wang, X.: Shunted self-attention via multi-scale token aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10853–10862 (2022)
14. Wang, W., et al.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
15. Chu, X., et al.: Twins: Revisiting the design of spatial attention in vision transformers. Advances in Neural Information Processing Systems, pp. 9355–9366 (2021)
16. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180 (2019)
17. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
18. Zhu, L., Wang, X., Ke, Z., Zhang, W., Lau, R. W.: BiFormer: Vision Transformer with Bi-Level Routing Attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10323–10333 (2023)
19. Huang, H., Zhou, X., Cao, J., He, R., Tan, T.: Vision Transformer with Super Token Sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10323–10333 (2023)
20. Grainger, R., Paniagua, T., Song, X., Cuntoor, N., Lee, M. W., Wu, T.: PaCa-ViT: Learning Patch-to-Cluster Attention in Vision Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22690–22699 (2023)
21. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021, LNCS, vol. 12901, pp. 14–24. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_2
22. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: FAT-Net: Feature adaptive transformers for automated skin lesion segmentation: Medical Image Anal. **76**, 102327 (2022)
23. Valanarasu, J. M., Oza, P., Hacihaliloglu, I., Patel, V. M.: Medical transformer: Gated axial-attention for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021, LNCS, vol. 12901, pp. 36–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_4
24. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imag. **37**(11), 2514–2525 (2018) multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data. **5**(1), 1–9 (2018)
25. Li, X., et al.: The state-of-the-art 3D anisotropic intracranial hemorrhage segmentation on non-contrast head CT: The INSTANCE challenge. arXiv preprint arXiv:2301.03281 (2023)
26. Zhou, H. Y., et al.: nnFormer: Volumetric medical image segmentation via a 3D transformer. IEEE Trans. Image Process. **42**, 4036–4045 (2023)

27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W. M., Frangi, A.F. (eds.) MICCAI 2015, LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
28. Gu, R., et al.: CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imag. **40**(2), 699–711 (2020)
29. Chen, G., Li, L., Dai, Y., Zhang, J., Yap, M. H.: AAU-net: an adaptive attention U-net for breast lesions segmentation in ultrasound images. IEEE Trans. Med. Imag. **42**(5), 1289–1300 (2023)
30. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., Maier-Hein, K. H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods. **18**(2), 2023–2011 (2021)
31. Chen, J., et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
32. He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2former: An efficient hierarchical hybrid transformer for medical image segmentation. IEEE Trans. Med. Imag. **42**(9), 2763–2775 (2023)
33. Roy, S., et al.: Mednext: transformer-driven scaling of convnets for medical image segmentation.. In: Greenspan, H., et al. (eds.) MICCAI 2023, LNCS, vol. 14223, pp. 405–415. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_39
34. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: Multimodal brain tumor segmentation using transformer. In: de Bruijne, M., et al. (eds.) MICCAI 2021, LNCS, vol. 12901, pp. 109–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_11