DTMFormer: Dynamic Token Merging for Boosting Transformer-Based Medical Image Segmentation

Zhehao Wang[1]*, Xian Lin[1]*, Nannan Wu[1], Li Yu[1], Kwang-Ting Cheng[2], Zengqiang Yan[1]†

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China
[2]School of Engineering, Hong Kong University of Science and Technology, Hong Kong
{zhehao_wang, xianlin, wnn2000, hustlyu}@hust.edu.cn, timcheng@ust.hk, z_yan@hust.edu.cn

Abstract

Despite the great potential in capturing long-range dependency, one rarely-explored underlying issue of transformer in medical image segmentation is attention collapse, making it often degenerate into a bypass module in CNN-Transformer hybrid architectures. This is due to the high computational complexity of vision transformers requiring extensive training data while well-annotated medical image data is relatively limited, resulting in poor convergence. In this paper, we propose a plug-n-play transformer block with dynamic token merging, named DTMFormer, to avoid building long-range dependency on redundant and duplicated tokens and thus pursue better convergence. Specifically, DTMFormer consists of an attention-guided token merging (ATM) module to adaptively cluster tokens into fewer semantic tokens based on feature and dependency similarity and a light token reconstruction module to fuse ordinary and semantic tokens. In this way, as self-attention in ATM is calculated based on fewer tokens, DTMFormer is of lower complexity and more friendly to converge. Extensive experiments on publicly-available datasets demonstrate the effectiveness of DTMFormer working as a plug-n-play module for simultaneous complexity reduction and performance improvement. We believe it will inspire future work on rethinking transformers in medical image segmentation. Code: https://github.com/iam-nacl/DTMFormer.

1 Introduction

Medical image segmentation is a fundamental task in computer-aided diagnosis, image-guided surgery, and treatment planning. Despite the great success of convolutional neural networks (CNN) (Milletari, Navab, and Ahmadi 2016; Ronneberger, Fischer, and Brox 2015; Zhou et al. 2018), its relatively limited receptive fields have been a major bottleneck, especially in dealing with small-size objects, irregular shapes, etc. Transformer (Vaswani et al. 2017; Dosovitskiy et al. 2021; Zheng et al. 2021), designed for capturing long-term dependencies and allowing the network to dynamically aggregate relevant features globally, seems a perfect complement for CNN. Therefore, developing CNN-Transformer hybrid architectures has been extensively studied for medical image segmentation.

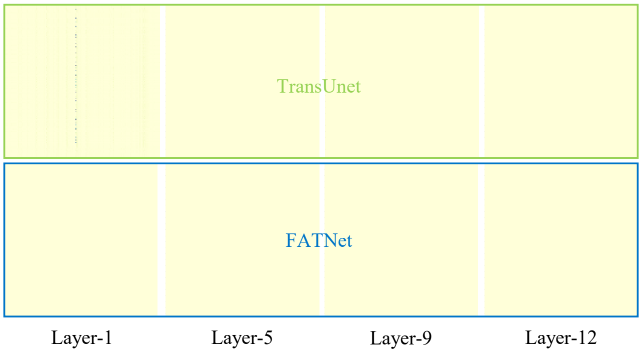

Figure 1: Attention collapse in existing transformer-based medical image segmentation approaches. The color of each token in a row represents its dependency score. The darker the color, the closer the dependency. All tokens sharing a uniform dependency distribution is called attention collapse.

Unfortunately, one underlying but rarely-explored issue in existing CNN-Transformer hybrid frameworks is attention collapse (Lin et al. 2023a) where all patches/tokens share the same dependency distribution. In other words, transformer becomes a bypass module, completely failing to extract meaningful global features, as illustrated in Fig. 1. The main reasons can be summarized as follows:

1. **High Model Complexity**. Vision transformer is of $O(n^2)$ model complexity where n is the length of input token sequences (Vaswani et al. 2017). In medical image segmentation, with the increase of patches/tokens, it becomes too complicated to well converge given relatively limited well-annotation training data, making transformers less effective in capturing long-range dependency.
2. **Severe Dependence Redundancy**. Medical images of the same modality share stable structures/views and in most cases rely on local features for segmentation. Consequently, building pair-wise dependence for all patches/tokens may produce severe dependence redundancy (Lin et al. 2023b), which in turn makes it more difficult to capture truly-useful long-range dependence.

Inspired by the above analysis, if we can filter out those redundant tokens in transformers and preserve those important tokens, it will not only reduce model complexity and

[*]Equal contribution
[†]Corresponding author

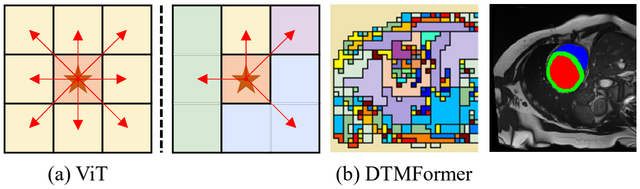

Figure 2: Illustration of dependence establishment in ViT and DTMFormer: (a) ViT builds pair-wise dependency for all tokens. (b) DTMFormer clusters vanilla tokens into fewer semantic tokens (*i.e.*, marked with different colors) and builds pair-wise dependency across semantic tokens.

dependence redundancy but also boost the convergence of transformers for long-range dependency establishment, thus leading to performance improvement. It should be noted that such a motivation is different from existing research on transformer pruning/sparsification whose goal is to pursue lightweight transformers with acceptable performance sacrifice. Comparatively, our primary goal is to pursue performance improvement instead of transformer de-complexity. To accomplish this, we propose a novel transformer block with dynamic token merging, named DTMFormer, to first cluster tokens into fewer semantic tokens and then build pair-wise dependence only on those semantic tokens, as illustrated in Fig. 2. Experiments on multiple widely-used datasets demonstrate the effectiveness of DTMFormer for not only complexity reduction but also stable performance improvement across various CNN-Transformer hybrid models. The main contributions are summarized as follows:

- A plug-n-play transformer block, DTMFormer, which can be inserted as a super-ordinate replacement into vanilla transformers for complexity reduction and performance improvement.
- An attention-guided token merging mechanism to address attention collapse in medical transformers as well as enhancing important tokens to accelerate convergence.
- A lightweight token reconstruction module to reconstruct and fuse cross-resolution tokens.

2 Related Work

2.1 Medical Image Segmentation

Medical image segmentation was mainly dominated by convolutional neural networks (CNN) (Milletari, Navab, and Ahmadi 2016; Ronneberger, Fischer, and Brox 2015; Mou et al. 2019; Peng et al. 2017; Zhou et al. 2018; Zhao et al. 2017), suffering from limited receptive fields. Recently, transformers (Vaswani et al. 2017; Dosovitskiy et al. 2021; Zheng et al. 2021), born to capture long-term interaction, have drawn extensive attention (Chen et al. 2021; Zhang, Liu, and Hu 2021; Valanarasu et al. 2021; Cao et al. 2022). Specifically, TransUNet (Chen et al. 2021) made the first attempt to combine transformers with U-Net to establish a self-attention mechanism, proving the potential of transformer working as a powerful encoder for medical image segmentation. TransFuse (Zhang, Liu, and Hu 2021) fused transformer and CNN in a parallel way to improve the efficiency of global environment modeling. To improve model efficiency, MedT (Valanarasu et al. 2021) introduced axial deeplab-based gated axial attention and a local-global training strategy. Swin-UNet (Cao et al. 2022) changed vanilla transformers into Swin transformers, and designed a symmetric Swin transformer-based decoder and a patch extension layer to perform up-sampling operations. MCTrans (Ji et al. 2021) incorporated rich contextual dependencies and semantic relations for accurate biomedical segmentation within a unified transformer network.

2.2 Token Sparsification in Transformer

Token Pruning Few works were proposed to learn token markers to identify and remove unimportant tokens for image classification like Power-BERT (Goyal et al. 2020), AViT(Yin et al. 2022), etc. On the one hand, most pruning operations are applied to the inference phase instead of the training phase, making them less flexible. On the other hand, though token pruning conveniently reduces model parameters, simply extending it to image segmentation may be counter-productive, resulting in severe under-segmentation.

Token Merging Among the few token merging approaches, SPViT (Kong et al. 2022) merged all unimportant and removable tokens as one token. TokenLearner (Ryoo et al. 2021) used a multilayer perceptron structure to reduce the number of tokens. Token pooling (Marin et al. 2021) used a K-means-based token merging method, but was not applicable to off-the-shelf models as its training is too slow. In addition, a token reduction strategy along with token fusion was proposed in TCFormer (Zeng et al. 2022), which is theoretically extendable to segmentation.

It should be noted that, till now, token sparsification is under-explored even in natural image processing, not to mention medical image segmentation in this work.

3 Methodology

3.1 Preliminaries

DTMFormer is to replace vanilla transformer blocks with lightweight self-attention calculation. Taking a transformer block with four transformer layers as an example as illustrated in Fig. 3. DTMFormer mainly consists of three modules: vanilla transformer layers, an Attention-guided Token Merging (ATM) module, and a Light Token Reconstruction (LTR) module, where ATM dynamically merges tokens into fewer semantic tokens and LTR reconstructs the original token resolution and fuse different types of tokens.

3.2 ATM

ATM is to merge redundant and similar tokens into fewer tokens with more semantic-aware information. It is formulated as clustering, including cluster center selection (*i.e.* determining semantic tokens) and clustering (*i.e.* assigning tokens to semantic tokens and updating semantic tokens).

Attention-guided Cluster Center Selection Selecting a suitable center token is crucial for token merging. Optimal center/semantic tokens should satisfy

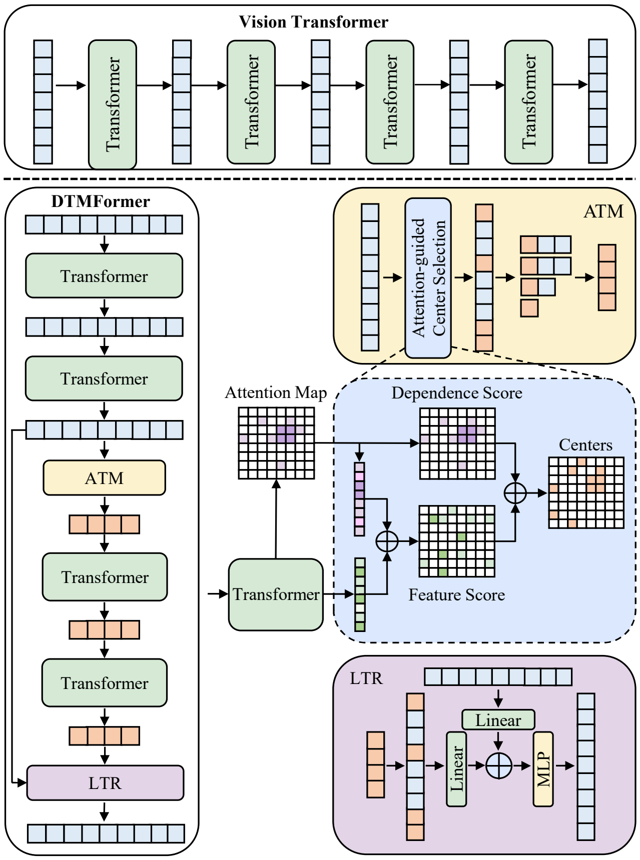

Figure 3: Overview of DTMFormer against ViT. Token sparsification is accomplished via an Attention-guided Token Merge (ATM) module to cluster tokens based on feature and dependency similarities and a Light Token Reconstruction (LTR) module to recover the original token resolution.

- majorly sampled from foreground and boundary regions, to guarantee segmentation performance.
- sparsely sampled from background and redundant regions, to avoid dependence redundancy.

To accomplish this, we propose to cluster tokens according to both dependence importance and feature similarity, as illustrated in Fig. 3. Feature similarity is based on the observation that medical images usually share stable structures/views, leading to lower intra-class variations. Therefore, tokens from the same category/organ are more likely to share similar features following related work (Zeng et al. 2022; Zhou et al. 2023). Based on feature similarity, it is expected to merge tokens from the background and large-scale foreground regions. Dependence importance is to fairly preserve tokens from boundaries and small-scale foreground regions, which is based on the observation of stable structures in medical images where pair-wise dependency across different organs/objects is more stable compared to natural images. In other words, tokens from the same category/organ are more likely to share similar dependency distributions. Therefore, dependency importance is chosen as one criterion for cluster center selection.

Given input feature map $X \in \mathbb{R}^{H\times W\times C}$, a sequence of tokens are generated as $I \in \mathbb{R}^{N\times C'}$, where $C' = p^2 \cdot C$ is the token channel determined by the patch size (p, p) and $N = HW/C$ is the total number of tokens. Then, given h-head self-attention and I, totally h attention maps $M \in \mathbb{R}^{h\times N\times N}$ are produced. To measure the dependence scores S_d of tokens, we column-wisely calculate the sum of M and sum up $\frac{h}{2}$ heads, defined as

$$S_d = \text{norm}\left(\sum_{i=1}^{\frac{h}{2}}\sum_{j=1}^{N} M(i,j,k)\right) \in \mathbb{R}^{N\times 1}. \tag{1}$$

Here, only half of the self-attention heads are used for calculation. It is to maintain the diversity of extracted features, as M will be penalized by an auxiliary loss during training (as discussed in Section 3.4).

In terms of the feature dimension, we directly apply DPC-KNN (Du, Ding, and Jia 2016) to tokens based on feature similarity. In optimal clustering, cluster centers are surrounded by low-density neighbors, and distant from other centers or high-density points. Inspired by this, we calculate two variables for each token i: the density ρ_i and the minimum distance to a higher density token δ_i.

As shown in Fig. 3, instead of directly using the token features produced by transformer layers, both dependence importance and token features are used. Specifically, we first compute the distance scores

$$D_{token}(i,j) = \frac{\|x_i - x_j\|_2^2}{\sqrt{C'}}, \tag{2}$$

and

$$D_{attn}(i,j) = \frac{\|y_i - y_j\|_2^2}{\sqrt{N}}, \tag{3}$$

where x_i, x_j represent any two tokens in I, and y_i, y_j are their dependence importance from S_d, $D_{token} \in \mathbb{R}^{N\times N}$ and $D_{attn} \in \mathbb{R}^{N\times N}$ represent the distance scores of all token pairs, and the final distance matrix D is defined as:

$$D = (1-\alpha)D_{token} + \alpha D_{attn}, \tag{4}$$

According to D, the local density ρ_i of token i to its K-nearest neighbors is calculated as

$$\rho_i = \exp\left(\sum_{j\in KNN(D(i))} D(i,j)\right), \tag{5}$$

where $KNN\left(D(i)\right)$ denotes the K-nearest neighbors of token i. Similarly, the minimum distance δ_i between token i and any other token with a higher density is obtained by

$$\delta_i = \begin{cases} \min_{j:\rho_j>\rho_i} D(i,j), \text{ if } \exists j \text{ s.t. } \rho_j > \rho_i \\ \max_j D(i,j), \text{ otherwise}. \end{cases} \tag{6}$$

To this end, the feature score of each token i is calculated as

$$\mathrm{S}_f^i = \rho_i\delta_i. \tag{7}$$

The final score S of all tokens is formulated as:

$$S = S_d + S_f, \tag{8}$$

based on which the top rN tokens (*i.e.*, r is the sparsity ratio) with the highest scores are selected as center/semantic tokens for merging.

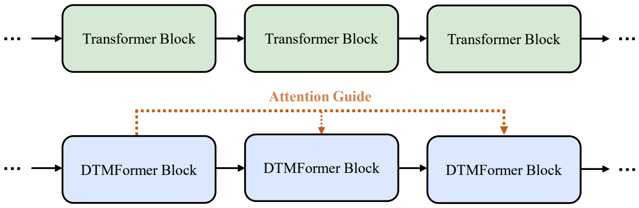

Figure 4: Illustration of how to stack ordinary ViT blocks and DTMFormer blocks respectively from the same scale.

Token Clustering/Merging Given center/semantic tokens, all other tokens are clustered based on feature similarity. Instead of directly averaging the tokens belonging to the same center, inspired by the attention mechanism in (Rao et al. 2021), we propose a token feature fusion method guided by the attention score, described as

$$z_i = \frac{\sum_{j \in C_i} e^{P_j} x_j}{\sum_{j \in C_i} e^{P_j}}, \tag{9}$$

where z_i represents the merged token features of each semantic token i, C_i indicates the set of tokens assigned to i, x_j represents the original token features in I, and P_j is the weight of token j learned based on x_j through linear projection layers.

In addition, for better feature fusion, the merged tokens are fed into a transformer layer as queries (*i.e.* Q), and the original tokens are projected into keys (*i.e.*, K) and values (*i.e.*, V) for self-attention calculation. To further up-weight important tokens, we add the weight scores P to the computation of attention, defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + P\right)V, \tag{10}$$

where $\sqrt{d_k}$ is the channel number of Q.

3.3 LTR

LTR is to aggregate two different scale tokens (*i.e.*, ordinary tokens and dynamically-merged semantic tokens) and reconstruct output tokens for the following down-sampling/up-sampling operations. As shown in Fig. 3, LTR only uses two linear layers and one MLP layer to reduce computation burdens. Specifically, for each merged token containing abstract semantics, LTR up-samples the token and recovers the feature mapping based on its merging history. During token merging in ATM, we record the positional correspondence between the original and merged tokens. In the up-sampling process of LTR, those records are to copy the merged token features into the corresponding up-sampled tokens.

3.4 Loss Functions

As described above, only the attention maps of the first module in DTMFormer are used for token merging (*i.e.*, dependence importance estimation) as illustrated in Fig. 4. To produce high-quality attention maps, we introduce an L1 loss onto S_d penalized by ground truth, denoted as L_{attn}. Thus, the overall loss is written as

$$L = L_{dice} + \lambda L_{CE} + \beta L_{attn}, \tag{11}$$

where L_{dice} represents the Dice loss, L_{CE} denotes the cross-entropy loss, λ and β are balancing hyper-parameters.

Why not Multi-Stage? Most existing work on transformer token sparsification used a multi-stage network structure (*i.e.*, gradually merging tokens). Comparatively, DTMFormer uses only a two-stage structure. On the one hand, such a two-stage structure is more flexible to work as a plug-and-play module. On the other hand, according to our experiments, using stages would bias more to abstract the semantics of tokens. In medical image segmentation, overly abstract semantics tokens would make tokens less distinguishable, resulting in attention collapse. Consequently, the later stages of such multi-stage structures often fail to work. Another major reason for not using multi-stage is to avoid error propagation. During token merging, each merged token will unavoidably contain irrelevant tokens (*e.g.*, background tokens). Given multi-stage clustering, merged tokens might be dominated by irrelevant tokens especially for small-scale organs/objects, leading to performance degradation.

4 Experiments

4.1 Datasets and Evaluation Metric

Three publicly-available and widely-used datasets are adopted for evaluation, including

1. ACDC: The ACDC dataset (Bernard et al. 2018) consists of short-axis cine-MRI from 150 patients with left ventricle (LV), right ventricle (RV), and myocardium (Myo) manually annotated by experienced experts on end-diastolic (ED) and end-systolic (ES) phase instants. In the official challenge, the ACDC dataset was divided into 100 and 50 patients for training and testing respectively. For a fair comparison, we strictly follow TransUNet (Chen et al. 2021) to divide the 100 patients into 7:1:2 for training, validation, and testing respectively.
2. ISIC: The ISIC dataset (Codella et al. 2019) contains a total of 2596 dermoscopic images with well-annotated labels. The same data split in APFormer (Lin et al. 2023b) is adopted for training and testing.
3. BTCV: The BTCV dataset is an abdominal multi-organ segmentation dataset, consisting of 50 abdominal CT scans with varying volume sizes (512×512×85-512×512×198 pixels), among which 30 scans with pixel-wise annotations of 13 organs are publicly available. Following the setting of most transformer-based methods (Cao et al. 2022; Xu et al. 2021; Huang et al. 2022; Lin et al. 2023b), the same train-test data split is adopted to select 18 CT scans for training and the rest 12 CT scans for testing, and eight of 13 organs are used for evaluation.

For a fair evaluation, the most commonly-used metrics are adopted, including Dice, Hausdorff Distance (HD), Intersection over Union (IoU), and Accuracy (Acc).

4.2 Experimental Setup

Implementation Details All methods were implemented within PyTorch and trained by the following settings:

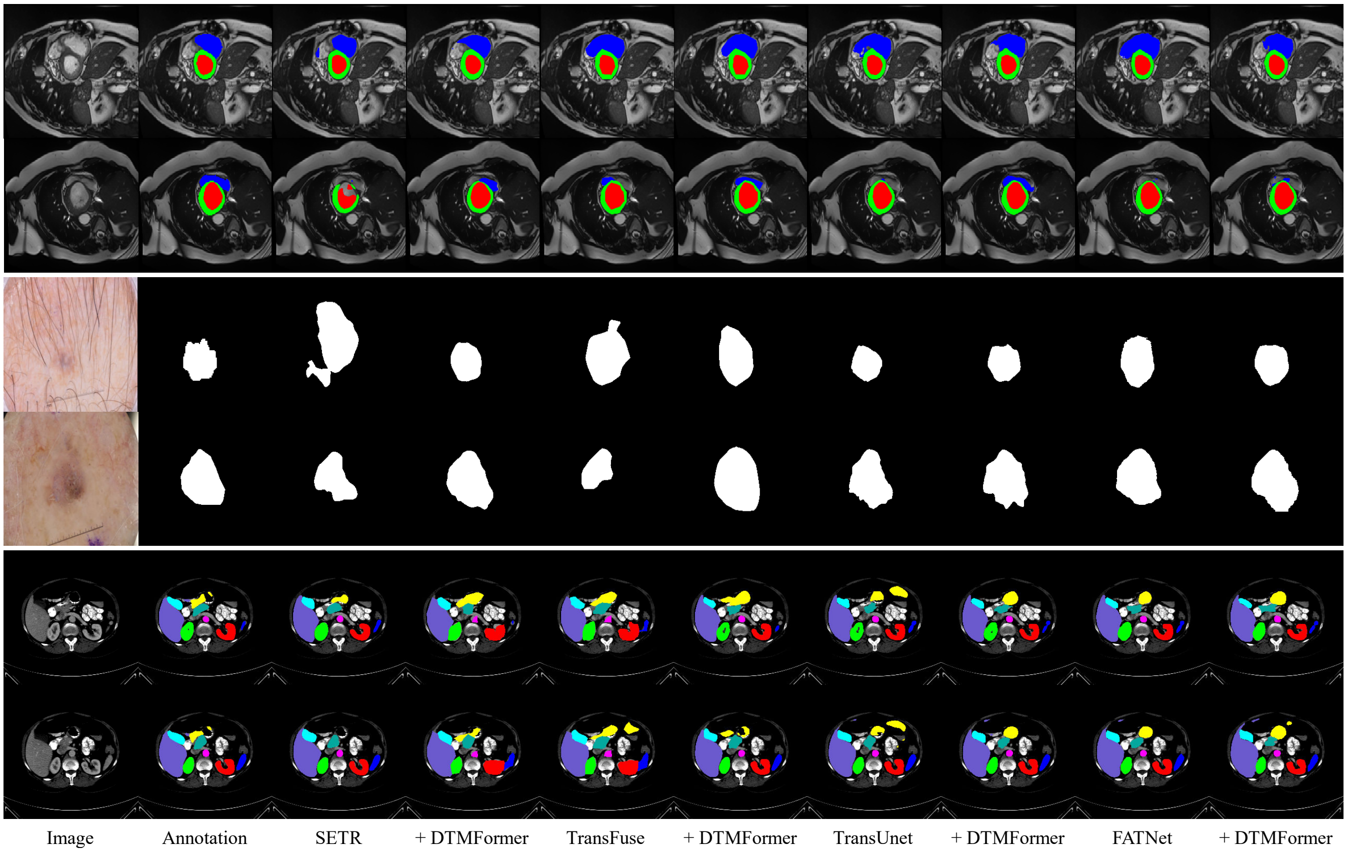

Figure 5: Qualitative comparison results with and without DTMFormer combined with various SOTA transformer-based architectures on the ACDC, ISIC, and BTCV datasets respectively. Different organs/targets are assigned with different colors.

- ACDC:bs=4; lr=5e-4; ep=600; opt=Adam;
- BTCV:bs=4; lr=1e-4; ep=500; opt=Adam;
- ISIC:bs=4; lr=5e-4; ep=400; opt=Adam;

For data augmentation. random rotation, random scaling, cropping, contrast adjustment, and gamma augmentation were applied. Particularly, for better convergence and comparison of TransFuse and TransFuse+DTMFormer, the initial learning rate is set as 5e-4 for BTCV.

Baseline Networks for Comparison. For a comprehensive evaluation, both pure transformer-based and CNN-Transformer hybrid architectures were selected to testify the plug-n-play nature of DTMFormer, including SETR (Zheng et al. 2021), TransFuse (Zhang, Liu, and Hu 2021), TransUnet (Chen et al. 2021), and FATNet (Wu et al. 2022).

4.3 Evaluation on ACDC

Comparison Methods. SOTA 2D and 3D task-specific approaches are adopted for evaluation, including nnFormer (Zhou et al. 2022), UNETR (Hatamizadeh et al. 2022), D-Former (Wu et al. 2023), SwinUnet (Cao et al. 2022), and MISSFormer (Huang et al. 2022).

Quantitative Comparison. As summarized in Table 1, among comparison approaches, D-Former achieves the best performance across all three organs. Without DTMFormer, all baseline networks are sub-optimal compared to SOTA comparison methods. By introducing DTMFormer, consistent performance improvements are achieved across all baseline networks, leading to an average increase of 0.47%-1.08% in Dice. More importantly, FATNet+DTMFormer achieves the best segmentation performance against SOTA comparison methods, outperforming D-Former.

Type	Method	Avg.	RV	Myo	LV
3D	nnFormer	92.06	90.94	89.58	95.65
	UNETR	88.61	85.29	86.52	94.02
	D-Former	92.29	91.33	**89.60**	95.93
2D	SwinUnet	90.00	88.55	85.62	95.83
	MISSFormer	90.86	89.55	88.04	94.99
2D	SETR	88.56	86.48	84.70	94.48
	+ DTMFormer	89.64	89.50	84.83	94.60
	TransFuse	90.48	90.43	86.43	94.58
	+ DTMFormer	90.95	90.70	87.21	94.93
	TransUnet	91.83	91.29	88.87	95.33
	+ DTMFormer	92.30	91.88	89.23	95.77
	FATNet	91.94	90.71	89.15	**95.95**
	+ DTMFormer	**92.53**	**92.21**	89.55	95.85

Table 1: Comparison measured in Dice on ACDC. The best and second-best results are bolded and underlined.

Type	Method	Dice	HD	IoU	Acc
CNN	Unet	88.15	7.93	80.64	95.48
	Att-UNet	88.53	7.79	81.44	95.69
	CPFNet	89.66	5.40	82.75	95.96
	CANet	89.81	5.11	83.12	96.23
	nnU-Net	90.22	-	83.76	**96.48**
	Ms RED	90.25	-	83.77	96.44
Hybrid /Trans	MedT	88.75	-	81.74	95.85
	Patcher	89.11	-	82.36	96.08
	SwinUNet	89.15	5.99	82.26	96.00
	MISSFormer	88.42	7.50	81.65	96.15
	LeViTUNet	89.85	5.45	83.39	96.28
	H2Former	89.45	6.02	82.74	96.10
Hybrid /Trans	SETR	88.42	6.39	81.07	96.02
	+ DTMFormer	88.98	6.77	82.15	95.95
	TransFuse	88.00	6.77	80.71	95.65
	+ DTMFormer	89.21	5.82	82.02	95.91
	TransUnet	89.19	5.73	82.12	96.04
	+ DTMFormer	89.67	5.76	82.90	96.04
	FATNet	89.05	6.73	81.86	95.95
	+ DTMFormer	**90.40**	**4.99**	**83.90**	96.38

Table 2: Comparison against SOTA methods on ISIC. The best and second-best results are bolded and underlined.

Qualitative Comparison. Qualitative comparison results of baseline networks with and without DTMFormer are illustrated in the first two rows of Fig. 5. Across various baseline networks, coupling DTMFormer effectively reduces false positives and false negatives, even for small-size organs, leading to better segmentation results.

4.4 Evaluation on ISIC

Comparison Methods. SOTA CNN-/Transformer-based task-specific approaches are adopted for evaluation, including UNet (Ronneberger, Fischer, and Brox 2015), Att-UNet (Schlemper et al. 2019), CPFNet (Feng et al. 2020), CANet (Gu et al. 2021), nnU-Net (Isensee et al. 2021), Ms RED (Dai et al. 2022), MedT (Valanarasu et al. 2021), and Patcher (Ou et al. 2022), SwinUNet (Cao et al. 2022), MISSFormer (Huang et al. 2022) ,LeVit-UNet-384 (Xu et al. 2021), and H2Former (He et al. 2023).

Quantitative Comparison. As summarized in Table 2, in general, CNN approaches are superior compared to both pure transformer-based and CNN-Transformer hybrid methods, indicating that transformers fail to enrich global features as expected. Through DTMForer to boost model convergence, all baseline networks benefit from meaningful long-range dependence exploration, achieving an average increase of 0.48%-1.35% in Dice. More importantly, FATNet+DTMFormer achieves the best performance across Dice, HD, and IoU and the second-best performance on Acc.

Qualitative Comparison. Qualitative comparison results of baseline networks with and without DTMFormer are illustrated in the middle two rows of Fig. 5. Across different cases, baseline networks can suffer from either over-segmentation or under-segmentation depending on the targets. By introducing DTMFormer, consistent improvements are achieved on all baseline networks, leading to better segmentation results.

4.5 Evaluation on BTCV

Comparison Methods. SOTA 2D CNN-/Transformer-based task-specific approaches are adopted for evaluation, including R50 U-Net (Chen et al. 2021), R50 Att-UNet (Chen et al. 2021), UNet (Ronneberger, Fischer, and Brox 2015), Att-UNet (Schlemper et al. 2019), SwinUNet (Cao et al. 2022), TransClaw U-Net (Yao et al. 2022), LeVit-UNet-384 (Xu et al. 2021), MT-UNet (Wang et al. 2022), MISSFormer (Huang et al. 2022), CA-GANformer (You et al. 2022), and APFormer (Lin et al. 2023b).

Quantitative Comparison. As summarized in Table 3, transformer-based approaches generally are superior compared to CNN-based methods, indicating the necessity of extracting global features. Among comparison methods, APFormer achieves the best performance of 83.53% in Dice. Comparatively, though the baseline network FATNet outperforms APFormer solely, introducing DTMFormer can further bring an average increase of 0.81% in Dice, leading to the best performance compared to all 2D approaches. More importantly, adopting DTMFormer achieves consistent performance improvements across all baseline networks, validating its flexibility to work as a plug-n-play module.

Qualitative Comparison. Qualitative comparison results of baseline networks with and without DTMFormer are illustrated in the last two rows of Fig. 5. Across various baseline networks, the main benefit of DTMFormer is reducing false positives, leading to "cleaner" segmentation maps.

4.6 Evaluation on Complexity Reduction

Though the primary goal of DTMFormer is to boost transformer convergence and segmentation performance, it is also expected to reduce model complexity. Quantitative results of baseline networks with and without DTMFormer measured in model parameters and GFLOPs are summarized in Table 4. As expected, introducing DTMFormer for dynamic token merging would effectively reduce model parameters and GFLOPs. As DTMFormer is only applied to transformer blocks, model efficiency improvements of baseline networks vary depending on how many transformer blocks are used.

4.7 Ablation Study

On the Hyper-parameter r. One key factor in DTMFormer is the hyper-parameter r that determines the sparsity of ATM for token merging. A series of ablation studies under various r are conducted on SETR as summarized in Table 5. In general, adopting any r would achieve performance improvements, validating the necessity of redundancy reduction in transformers for medical image segmentation. Gradually increasing r is more beneficial as it would improve the convergence property of transformers on fewer tokens. Further increasing r may be harmful, as important tokens might be wrongly merged, resulting in performance degradation.

Comparison with SOTA Token Sparsification Methods. As discussed above, token sparsification is under-explored even in natural image processing. To better validate the effectiveness of DTMFormer, both DViT (Rao et al. 2021) and CTM (Zeng et al. 2022) are re-implemented and included for

Type	Method	Avg.	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
CNN	R50 U-Net	74.68	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
	R50 Att-UNet	75.57	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
	U-Net	76.85	89.07	**69.72**	77.77	68.60	93.43	53.98	86.67	75.58
	Att-UNet	77.77	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
Hybrid /Trans	Swin-Unet	79.13	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
	TransClaw U-Net	78.09	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
	LeVit-Unet-384	78.53	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76
	MT-UNet	78.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
	MISSFormer	81.96	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
	CA-GANformer	82.55	89.05	67.48	86.05	82.17	95.61	67.49	91.00	81.55
	APFormer	83.53	90.84	64.36	90.54	85.99	94.93	72.16	91.88	77.55
Hybrid /Trans	SETR	75.23	84.91	53.70	81.01	76.60	94.55	52.17	89.98	68.89
	+ DTMFormer	79.31	86.20	53.49	87.74	82.23	95.25	61.32	89.96	78.30
	TransFuse	71.84	73.55	49.48	79.64	74.89	92.35	46.01	88.54	70.28
	+ DTMFormer	72.56	74.67	48.68	80.37	76.78	92.46	48.58	87.43	71.52
	TransUnet	81.99	89.23	59.05	88.25	85.19	95.79	69.02	91.43	77.97
	+ DTMFormer	82.27	90.49	58.56	90.04	85.38	95.83	67.09	91.61	79.12
	FATNet	84.44	91.28	63.58	91.13	**86.17**	**96.22**	73.92	92.39	80.85
	+ DTMFormer	**85.25**	**91.48**	65.02	**92.03**	85.37	95.98	**74.18**	**94.23**	**83.71**

Table 3: Comparison measured in Dice against SOTA 2D methods on BTCV. The best results are bolded.

Baseline	P (M)	Δ P	GFLOPS	Δ GFLOPs
SETR	51.5	↓ 18.0	53.2	↓ 43.4
+ DTMFormer	33.5		9.8	
TransFuse	143.0	↓ 9.7	82.8	↓ 17.9
+ DTMFormer	133.3		64.9	
TransUnet	93.2	↓ 9.3	32.3	↓ 17.8
+ DTMFormer	83.9		14.5	
FATNet	34.9	↓ 6.0	57.3	↓ 2.5
+ DTMFormer	28.9		54.8	

Table 4: Computational complexity analysis of DTMFormer on different baseline architectures.

r	Avg.	RV	Myo	LV	P (M)	GFLOPs
baseline	88.56	86.48	84.70	94.48	51.5	53.2
$1/2$	89.08	88.92	84.23	94.08	33.5	17.1
$1/4$	89.28	88.79	84.65	94.39	33.5	12.2
$1/8$	**89.64**	**89.50**	84.83	94.60	**33.5**	9.8
$1/16$	89.53	89.11	**84.87**	**94.61**	33.5	8.6
$1/32$	89.19	88.59	84.53	94.46	33.5	**8.0**

Table 5: Ablation study on the parameter r based on SETR evaluated on ACDC. The best results are bolded.

Method	Avg.	RV	Myo	LV	P (M)	GFLOPs
baseline	88.56	86.48	84.70	94.48	51.5	53.2
+ DViT	87.54	85.04	83.54	94.04	51.5	45.4
+ CTM	88.59	87.95	83.60	94.25	20.9	**3.4**
+ ATM	**89.64**	**89.50**	**84.83**	**94.60**	**33.5**	9.8

Table 6: Comparison with token sparsification methods on SETR evaluated on ACDC. The best results are bolded.

k	Avg.	RV	Myo	LV
5	89.64	**89.50**	84.83	94.60
10	**89.67**	89.45	**84.93**	**94.62**
20	89.59	89.45	84.72	94.61
100	89.40	89.30	84.52	94.38

Table 7: Ablation study of k on ACDC measured in Dice. The best results are bolded.

comparison as summarized in Table 6. DViT can hardly improve model efficiency while encountering significant performance degradation. Comparatively, CTM effectively reduces model complexity but fails to boost segmentation performance. Comparatively, introducing ATM would outperform the baseline with large margins and better model efficiency. It should be noted that such model efficiency improvements are acceptable as the primary goal of DTMFormer is for performance improvement.

On the Hyper-parameter k. In DTMFormer, token clustering is based on both dependency importance and feature similarity and k-NN is only used for density calculation in measuring feature similarity. Thus, varying k will not significantly affect the overall performance of DTMFormer. As summarized in Table 7, with the increase of k, density calculation may be dominated by large-size objects, resulting in wrongly-merged tokens of small-scale objects and performance degradation. Additionally, using a larger k will bring higher computational complexity. Thus, k shall be tuned from a small value to balance performance and complexity.

5 Conclusion

In this paper, we propose to boost the segmentation performance of transformers in medical image segmentation through dynamic token merging and propose DTMFormer working as a plug-n-play module for model efficiency and segmentation performance improvement. By clustering tokens into fewer semantic tokens, DTMFormer has a better convergence property, thus addressing attention collapse. Experiments on multiple publicly available datasets demonstrate the superiority of DTMFormer in achieving consistent performance improvement across various baseline networks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62271220 and 62202179 and in part by the Natural Science Foundation of Hubei Province of China under Grant 2022CFB585.

References

Bernard, O.; et al. 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans. Med. Imaging*, 37(11): 2514–2525.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV*, 205–218.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. TransUNet: Transformers make strong encoders for medical image segmentation. arxiv:2102.04306.

Dai, D.; Dong, C.; Xu, S.; Yan, Q.; Li, Z.; Zhang, C.; and Luo, N. 2022. Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. *Med. Image Anal.*, 75: 102293.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Du, M.; Ding, S.; and Jia, H. 2016. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowl. Based. Syst.*, 99: 135–145.

Feng, S.; Zhao, H.; Shi, F.; Cheng, X.; Wang, M.; Ma, Y.; Xiang, D.; Zhu, W.; and Chen, X. 2020. CPFNet: Context pyramid fusion network for medical image segmentation. *IEEE Trans. Med. Imaging*, 39(10): 3008–3018.

Goyal, S.; Choudhury, A. R.; Raje, S.; Chakaravarthy, V.; Sabharwal, Y.; and Verma, A. 2020. PoWER-BERT: Accelerating bert inference via progressive word-vector elimination. In *ICML*, 3690–3699.

Gu, R.; Wang, G.; Song, T.; Huang, R.; Aertsen, M.; Deprest, J.; Ourselin, S.; Vercauteren, T.; and Zhang, S. 2021. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans. Med. Imaging*, 40(2): 699–711.

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022. UNETR: Transformers for 3D medical image segmentation. In *WACV*, 574–584.

He, A.; Wang, K.; Li, T.; Du, C.; Xia, S.; and Fu, H. 2023. H2Former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Trans. Med. Imaging*, 42(9): 2763–2775.

Huang, X.; Deng, Z.; Li, D.; Yuan, X.; and Fu, Y. 2022. Missformer: An effective transformer for 2D medical image segmentation. *IEEE Trans. Med. Imaging*, 42(5): 1484–1494.

Isensee, F.; Jaeger, P. F.; Kohl, S. A. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods*, 18(2): 203–211.

Ji, Y.; Zhang, R.; Wang, H.; Li, Z.; Wu, L.; Zhang, S.; and Luo, P. 2021. Multi-compound transformer for accurate biomedical image segmentation. In *MICCAI*, 326–336.

Kong, Z.; Dong, P.; Ma, X.; Meng, X.; Sun, M.; Niu, W.; Shen, X.; Yuan, G.; Ren, B.; Qin, M.; Tang, H.; and Wang, Y. 2022. SPViT: Enabling faster vision transformers via soft token pruning. arxiv:2112.13890.

Lin, X.; Yan, Z.; Deng, X.; Zheng, C.; and Yu, L. 2023a. ConvFormer: Plug-and-play cnn-style transformers for improving medical image segmentation. In *MICCAI*, 642–651.

Lin, X.; Yu, L.; Cheng, K.-T.; and Yan, Z. 2023b. The lighter the better: Rethinking transformers in medical image segmentation through adaptive pruning. *IEEE Trans. Med. Imaging*, 42(8): 2325–2337.

Marin, D.; Chang, J.-H. R.; Ranjan, A.; Prabhu, A.; Rastegari, M.; and Tuzel, O. 2021. Token pooling in vision transformers. arxiv:2110.03860.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 565–571.

Mou, L.; Zhao, Y.; Chen, L.; Cheng, J.; Gu, Z.; Hao, H.; Qi, H.; Zheng, Y.; Frangi, A.; and Liu, J. 2019. CS-Net: Channel and spatial attention network for curvilinear structure segmentation. In *MICCAI*, 721–730.

Ou, Y.; Yuan, Y.; Huang, X.; Wong, S. T. C.; Volpi, J.; Wang, J. Z.; and Wong, K. 2022. Patcher: Patch transformers with mixture of experts for precise medical image segmentation. arXiv:2206.01741.

Peng, C.; Zhang, X.; Yu, G.; Luo, G.; and Sun, J. 2017. Large kernel matters – Improve semantic segmentation by global convolutional network. In *CVPR*, 1743–1751.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C.-J. 2021. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 13937–13949.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241.

Ryoo, M.; Piergiovanni, A.; Arnab, A.; Dehghani, M.; and Angelova, A. 2021. TokenLearner: Adaptive space-time tokenization for videos. In *NeurIPS*, 12786–12797.

Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; and Rueckert, D. 2019. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.*, 53: 197–207.

Valanarasu, J. M. J.; Oza, P.; Hacihaliloglu, I.; and Patel, V. M. 2021. Medical transformer: Gated axial-attention for medical image segmentation. In *MICCAI*, 36–46.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wang, H.; Xie, S.; Lin, L.; Iwamoto, Y.; Han, X.-H.; Chen, Y.-W.; and Tong, R. 2022. Mixed transformer U-Net for medical image segmentation. In *ICASSP*, 2390–2394.

Wu, H.; Chen, S.; Chen, G.; Wang, W.; Lei, B.; and Wen, Z. 2022. FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. *Med. Image Anal.*, 76: 102327.

Wu, Y.; Liao, K.; Chen, J.; Wang, J.; Chen, D. Z.; Gao, H.; and Wu, J. 2023. D-former: A U-shaped dilated transformer for 3D medical image segmentation. 35(2): 1931–1944.

Xu, G.; Wu, X.; Zhang, X.; and He, X. 2021. LeViT-UNet: Make faster encoders with transformer for medical image segmentation. arXiv:2107.08623.

Yao, C.; Hu, M.; Li, Q.; Zhai, G.; and Zhang, X.-P. 2022. Transclaw U-Net: Claw U-Net with transformers for medical image segmentation. In *ICICSP*, 280–284.

Yin, H.; Vahdat, A.; Alvarez, J. M.; Mallya, A.; Kautz, J.; and Molchanov, P. 2022. A-ViT: Adaptive tokens for efficient vision transformer. In *CVPR*, 10809–10818.

You, C.; Zhao, R.; Liu, F.; Chinchali, S. P.; Topcu, U.; Staib, L.; and Duncan, J. 2022. Class-aware generative adversarial transformers for medical image segmentation. arXiv:2201.10737.

Zeng, W.; Jin, S.; Liu, W.; Qian, C.; Luo, P.; Ouyang, W.; and Wang, X. 2022. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *CVPR*, 11101–11111.

Zhang, Y.; Liu, H.; and Hu, Q. 2021. TransFuse: Fusing transformers and cnns for medical image segmentation. In *MICCAI*, 14–24.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*, 6230–6239.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H. S.; and Zhang, L. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. arxiv:2012.15840.

Zhou, H.-Y.; Guo, J.; Zhang, Y.; Yu, L.; Wang, L.; and Yu, Y. 2022. nnFormer: Interleaved transformer for volumetric segmentation. arXiv:2109.03201.

Zhou, Y.; Zhu, H.; Liu, Q.; Chang, S.; and Guo, M. 2023. MonoATT: Online monocular 3D object detection with adaptive token transformer. In *CVPR*, 17493–17503.

Zhou, Z.; Rahman Siddiquee, M. M.; Tajbakhsh, N.; and Liang, J. 2018. UNet++: A nested U-Net architecture for medical image segmentation. In *DLMIA*, 3–11.