Saliency-Driven Token Merging for Vision Transformers

Supplementary Material

Algorithm 1 Overall Procedure of SAD-TM

Input: Pretrained L-layer model Θ; Merging rate m_r; Inference dataset $\mathcal{D}$; Batch size B; Weight hyperparameter α and β.

Calculate saliency outlier score according to pretrained model Θ ▷ According to Equations (5) to (7)
Inference Stage:
for each batch $b \subseteq \mathcal{D}$ **do**
  for each layer $l = 1, 2, ..., L$ **do**
    Calculate class attention in forward propagation ▷ According to Equation (8)
    Calculate class attention importance score $\mathbf{a}'_l$ ▷ According to Equation (9)
    Obtain final importance score with α and β ▷ According to Equation (10)
    Sort all tokens by $\mathbf{m}_l$ in descending order
    Divide token set into $\mathcal{C}$ and $\mathcal{M}$ according to m_r
    for each token $\mathbf{t}'_j \in \mathcal{M}$ **do**
      Select $\mathbf{t}'_i \in \mathcal{C}$ with $\arg\max_{\mathbf{t}'_i} \text{sim}(\mathbf{t}'_i, \mathbf{t}'_j)$
      Merge $\mathbf{t}'_j$ with $\mathbf{t}'_i$ ▷ According to Equation (12)
    end for
  end for
end for

A. Algorithm Diagram of SAD-TM

As mentioned in the main text, we present a concise algorithmic flowchart of SAD-TM here, as shown in Algorithm 1. The flowchart demonstrates the efficiency of our method, which eliminates the need for any additional training or fine-tuning. The saliency outlier scores can be obtained through just one complete forward and backward pass, and this process does not involve any modification of model parameters.

B. Additional Experimental Results

As described in Section 4.2 of the main text, we conduct experiments on extra ViT architectures. The three ViTs are DeiT-Base, MAE-Base, and MAE-Large, and their results are shown in Tables 7 to 9, respectively. On the DeiT-Base model, SAD-TM and SAD-TM-DM maintain a Top-1 Accuracy of 80.60% and 81.05% while reducing FLOPs by 42.61% and 40.34%. With a lower FLOPs compression rate, SAD-TM-DM still outperforms other methods such as IA-RED^2, EViT, and ToMe. Particularly, when FLOPs are reduced by approximately 28%, the accuracy of SAD-TM-DM (81.75%) is close to the baseline (81.81%). On MAE-Base, SAD-TM achieves an accuracy of 82.32% with a 37.50% reduction in FLOPs, which surpasses ToMe and EViT. On the MAE-Large model, SAD-TM-DM achieves an accuracy of 85.37% with a 38.31% reduction in FLOPs, outperforming EViT. Overall, SAD-TM and SAD-TM-DM can achieve significant FLOPs reduction under the three models while maintaining high accuracy, and the results demonstrate effectiveness and robustness in balancing model efficiency and performance.

Table 7. Results with DeiT-Base model. The results displayed in bold represent the best under specific FLOPs.

Method	Acc@1(%)	FLOPs(G)	FLOPs Redu.(%)
Baseline	81.81	17.6	-
IA-RED^{2*}	80.30	11.8	32.95
EViT*	81.30	11.6	34.09
DTEM*	81.01	11.6	34.09
EViT	80.37	11.5	34.66
ToMe	80.58	11.5	34.66
SAD-TM	81.13	11.5	34.66
SAD-TM-DM	**81.75**	**12.6**	**28.41**
SAD-TM-DM	**81.36**	**11.5**	**34.66**
SAD-TM-DM	**81.25**	**11.3**	**35.80**
POWER*	80.10	10.4	40.91
SCOP*	79.70	10.2	42.05
SAD-TM-DM	**81.05**	**10.5**	**40.34**
SAD-TM	**80.60**	**10.1**	**42.61**

Table 8. Results with MAE-Base model. The results displayed in bold represent the best under specific FLOPs.

Method	Acc@1(%)	FLOPs(G)	FLOPs Redu.(%)
Baseline	83.68	17.6	-
ToMe	82.32	11.5	34.66
EViT	82.01	11.5	34.66
SAD-TM	**82.46**	**11.5**	**34.66**
SAD-TM	**82.32**	**11.0**	**37.50**

C. Additional Results on MSE

As described in Section 3.3 of the main text, we conduct additional experiments on the model regarding the MSE between blocks of the model. In particular, we extend our investigation to include DeiT-Base and MAE-Large archi-

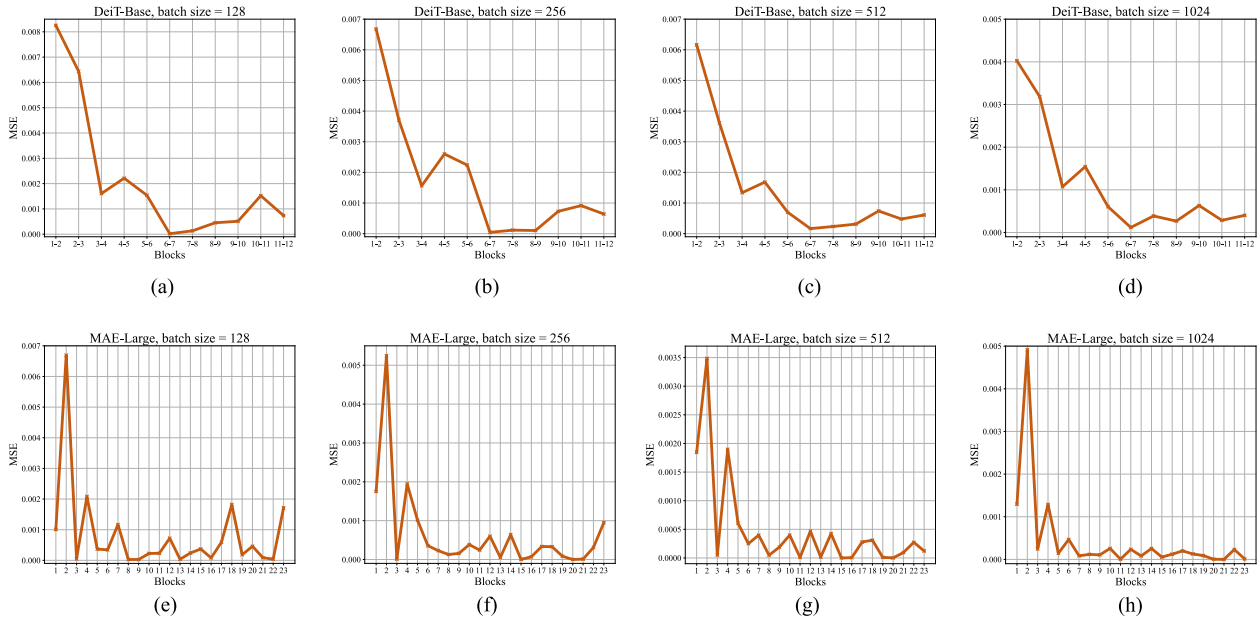

Figure 5. MSE under different batch sizes with DeiT-Base ((a), (b), (c), and (d)) and MAE-Large ((e), (f), (g), and (h)) models.

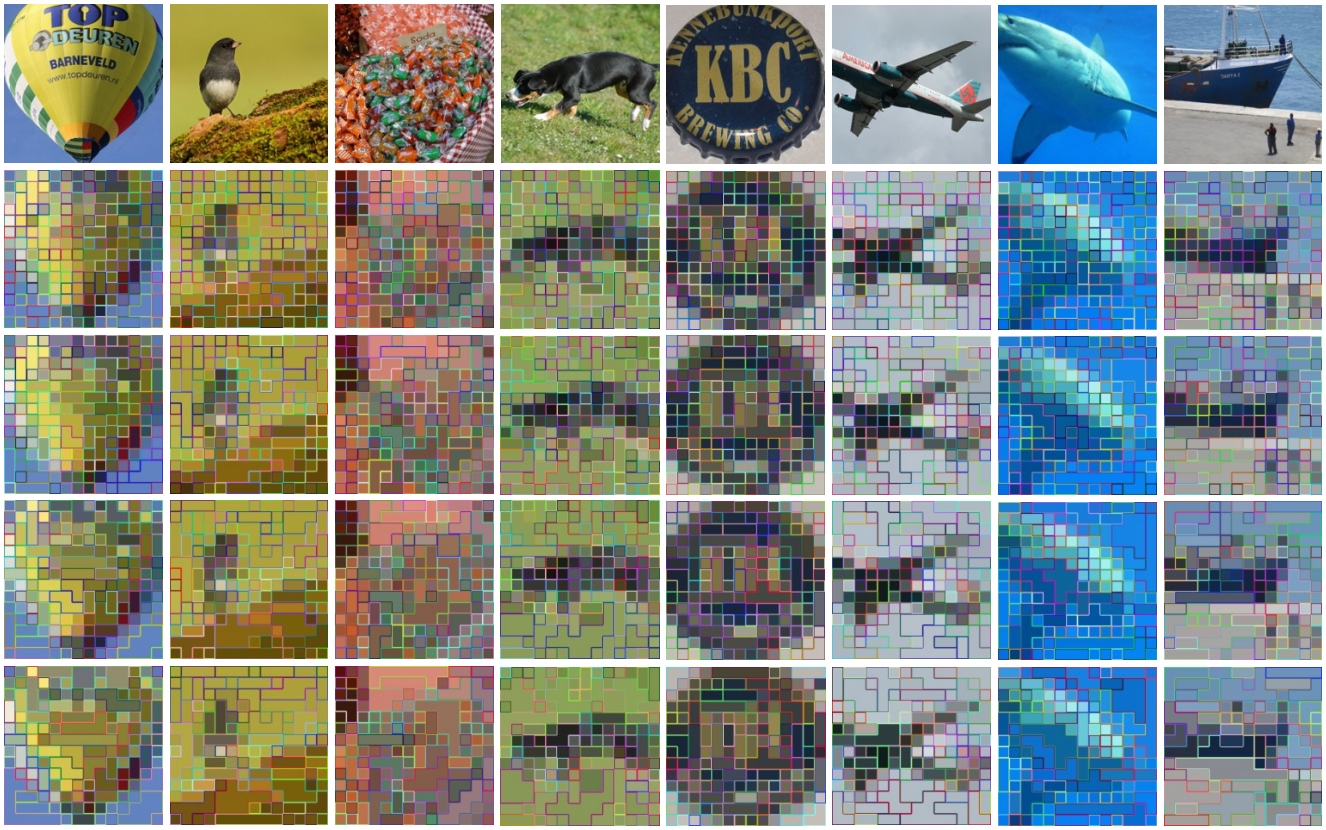

Figure 6. Visualization results of our SAD-TM on DeiT-Tiny model under the FLOPs of 0.8G. The images in each row represent the original image and the corresponding intermediate merging results after the third block, the sixth block, the ninth block, and the final merging results, respectively.

Figure 7. Visualization results of our SAD-TM on DeiT-Base model under the FLOPs of 10.1G. The images in each row represent the original image and the corresponding intermediate merging results after the third block, the sixth block, the ninth block, and the final merging results, respectively.

tectures, testing them under varying batch sizes. The corresponding results are illustrated in Figure 5. Our experimental findings reveal that the trends in MSE for both models align closely with those previously observed in the main text. This consistency across architectures underscores the generalizability of our observations and reinforces the validity of our findings across a diverse set of pretrained ViTs. Moreover, these results provide empirical support for the robust performance of our proposed variant, SAD-TM-DM, which remains effective under varying ViT architectures and a range of model compression configurations.

D. Extra Visualization Results

As shown in Figures 6 and 7, we visualize the intermediate and final merging results of SAD-TM on DeiT-Tiny and DeiT-Base models. The two Figures show, in the first row, randomly selected images from the preprocessed ImageNet validation set, and in the second, third, fourth, and fifth rows, the merging results after the third, sixth, and ninth blocks, as well as the final merged results of the model, respectively. We believe that presenting the visualized intermediate and final merging results will be more conducive to providing a more intuitive understanding and explanation on the merging process of our token merging method.

Table 9. Results with MAE-Large model. The results displayed in bold represent the best under specific FLOPs.

Method	Acc@1(%)	FLOPs(G)	FLOPs Redu.(%)
Baseline	85.98	61.6	-
EViT	85.06	39.6	35.71
SAD-TM	85.28	40.4	34.42
SAD-TM-DM	**85.37**	**38.0**	**38.31**
SAD-TM-DM	**85.34**	**37.5**	**39.12**
SAD-TM-DM	**85.24**	**37.0**	**39.94**

E. Results on Cross-Model Saliency Ablation

As described in Section 4.3.2 of the main text, we conduct cross-model saliency ablation experiments on six different models: DeiT-Tiny, DeiT-Small, DeiT-Base, MAE-Base, MAE-Large, and LV-ViT-S. The results are shown in Figure 8. It can be seen from the Figure that, even across

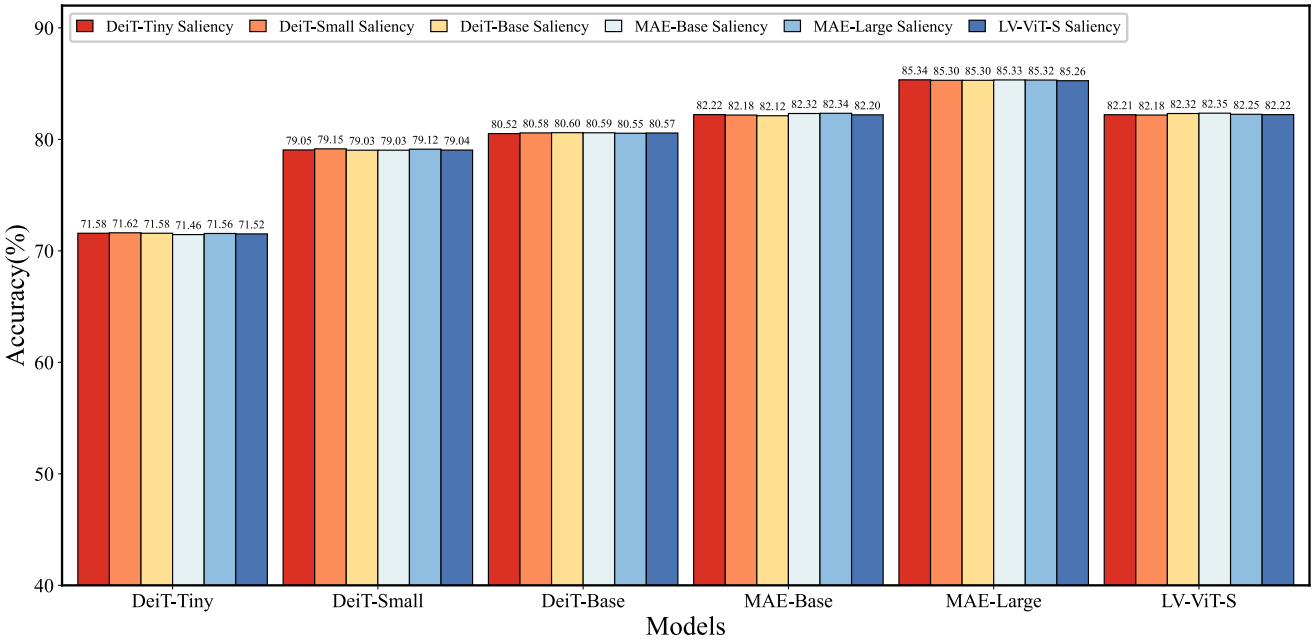

Figure 8. Cross-model ablation results. The FLOPs of the six models are 0.8G, 2.5G, 10.1G, 11.0G, 40.4G, and 3.6G, respectively.

different ViT architectures, our saliency-based token merging criterion remains universally applicable with negligible performance degradation (maximum-to-minimum variation not exceeding 0.2% across each experimental set). This further enhances the efficiency of SAD-TM by enabling token merging for all models to be guided by saliency outlier scores obtained from any ViT architecture.