Supplementary Material for Multi-criteria Token Fusion with One Step Ahead-Attention for Efficient Vision Transformers

A. Implementation details

For a comparison with previous works, we first evaluate MCTF with DeiT [7] on ImageNet-1K [2]. Following [4, 6, 9], we finetune the model with the pre-trained weights for 30 epochs with the batch size of 1,024 under 8 RTX3090 GPUs. We opt for the least epochs among previous works (*e.g.*, 30 for DynamicViT [6], 60 for SPViT [4], 100 for A-ViT [9]). For finetuning, the learning rate is initially set to 3e-5 and decreases to 1e-6 by the cosine annealing [5] with a cooldown of 10 epochs. Also, we finetune the T2T-ViT [10] and LV-ViT [3] with the initial learning rate of 5e-6, 1e-5 decreasing to 5e-7, 2e-6 for 30 epochs followed by 10 cooldown epochs, respectively. We do not use mixup-based augmentation [11, 12] to prevent the corrupted representation in fused tokens caused by the token fusion between different samples. Since we already track the size of the tokens, we also adopt proportional attention of ToMe [1] which simply update the attention scores with the size of the tokens $\mathbf{s}$ as $\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{C}} + \log \mathbf{s}\right)$. Regarding hyper-parameters for MCTF, we use $[\tau^{\text{info}}, \tau^{\text{sim}}, \tau^{\text{size}}] = [1, 1/20, 1/40]$ for the temperature parameters. And, We opt $\lambda = 1$ for DeiT-T and T2T-ViT, and $\lambda = 3$ for DeiT-S and LV-ViT for the coefficient of consistency loss. Similar to UDA [8], the consistency loss is calculated only with the sample that has a confidence score higher than $\beta = 0.4$. We also set the safeguard for excessive fusion by maintaining at least 10 tokens. For measuring the efficiency, we use `fvcore` and report the FLOPs of the model.

B. Analyses on MCTF

B.1. Sensitivity analysis on hyper-parameters of MCTF

To analyze the sensitivity of the hyper-parameters in MCTF, we compare the accuracy according to the temperature parameter τ in Table A. While evaluating each parameter, other hyper-parameters are set to default values mentioned in the implementation details of the main paper. We run the experiments with DeiT-S equipped with MCTF ($r = 16$). The default settings for each hyper-parameter are highlighted.

Table A. **Sensitivity analysis on the hyper-parameters.**

τ_{sim}	1	1/5	1/10	1/20	1/40	1/100
acc.	80.1	79.6	79.2	78.6	78.1	77.5

τ_{info}	1	1/5	1/10	1/20	1/40	1/100
acc.	78.7	79.8	80.0	80.1	80.0	79.8

τ_{size}	1	1/5	1/10	1/20	1/40	1/100
acc.	79.5	79.8	80.0	80.0	80.1	80.0

B.2. Loss of information

In this subsection, we measure the loss of information to validate the efficacy of MCTF. For this, we consider the cosine similarity between the class tokens with and without MCTF ($r = 16$) as a metric to measure the loss of information, which indicates the changes in the class tokens. In other words, if the similarity between class tokens is low, we infer that the fused tokens significantly affect the class token's representation while losing the information of original contents. The differences between the class tokens at each block are reported in Table B. As shown in the table, at the early stage of the Transformer

Table B. **Cosine similarity between the class tokens with and without MCTF per block.**

Criteria			Block index											
$\mathbf{W}^{\text{sim}}$	$\mathbf{W}^{\text{info}}$	$\mathbf{W}^{\text{size}}$	1	2	3	4	5	6	7	8	9	10	11	12
✓			1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9988	0.9973	0.9933	0.9870	0.9837	0.9695
	✓		1.0000	1.0000	0.9999	0.9996	0.9992	0.9976	0.9939	0.9887	0.9750	0.9550	0.9470	0.9153
		✓	1.0000	1.0000	0.9998	0.9996	0.9991	0.9968	0.9913	0.9812	0.9575	0.9141	0.9040	0.8546
✓	✓		1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9982	0.9958	0.9925	0.9907	0.9833
✓	✓	✓	1.0000	1.0000	1.0000	1.0000	0.9999	**0.9997**	**0.9992**	**0.9984**	**0.9961**	**0.9929**	**0.9914**	**0.9844**

(*e.g.*, [1-6]-th block), there is no big gap among the diverse criteria. However, as the number of fused tokens increases through consecutive blocks, there are substantial changes in the class tokens. Specifically, when we consider a single criterion, similarity is the best option for mitigating the loss of information compared to informativeness and size. Then, adopting the dual criterion composed of similarity and informativeness, we further lessen the changes between the class tokens showing the high similarity even in the rear block (*e.g.*, [7-12]-th block). At last, MCTF with all three criteria shows better similarity than dual-criteria. We believe that this minimization of information loss by adopting multi-criteria leads to consistent improvements compared to other single and dual criteria in image classification.

B.3. Qualitative comparison for one-step-ahead attention

In MCTF, the attention map $\hat{\mathbf{A}}^{l+1}$ of the fused tokens $\hat{\mathbf{X}}^{l}$ is approximated by aggregating the one-step-ahead attention $\mathbf{A}^{l+1}$, which is the attention before token fusion. The main paper shows that this approximation brings substantial speed improvements without any performance degradation by avoiding the re-computation of self-attention. In parallel, we here provide a qualitative comparison to show the soundness of our approaches. The visualization of the attention map in the [3,6,9,12]-th layer is provided in Figure 1.

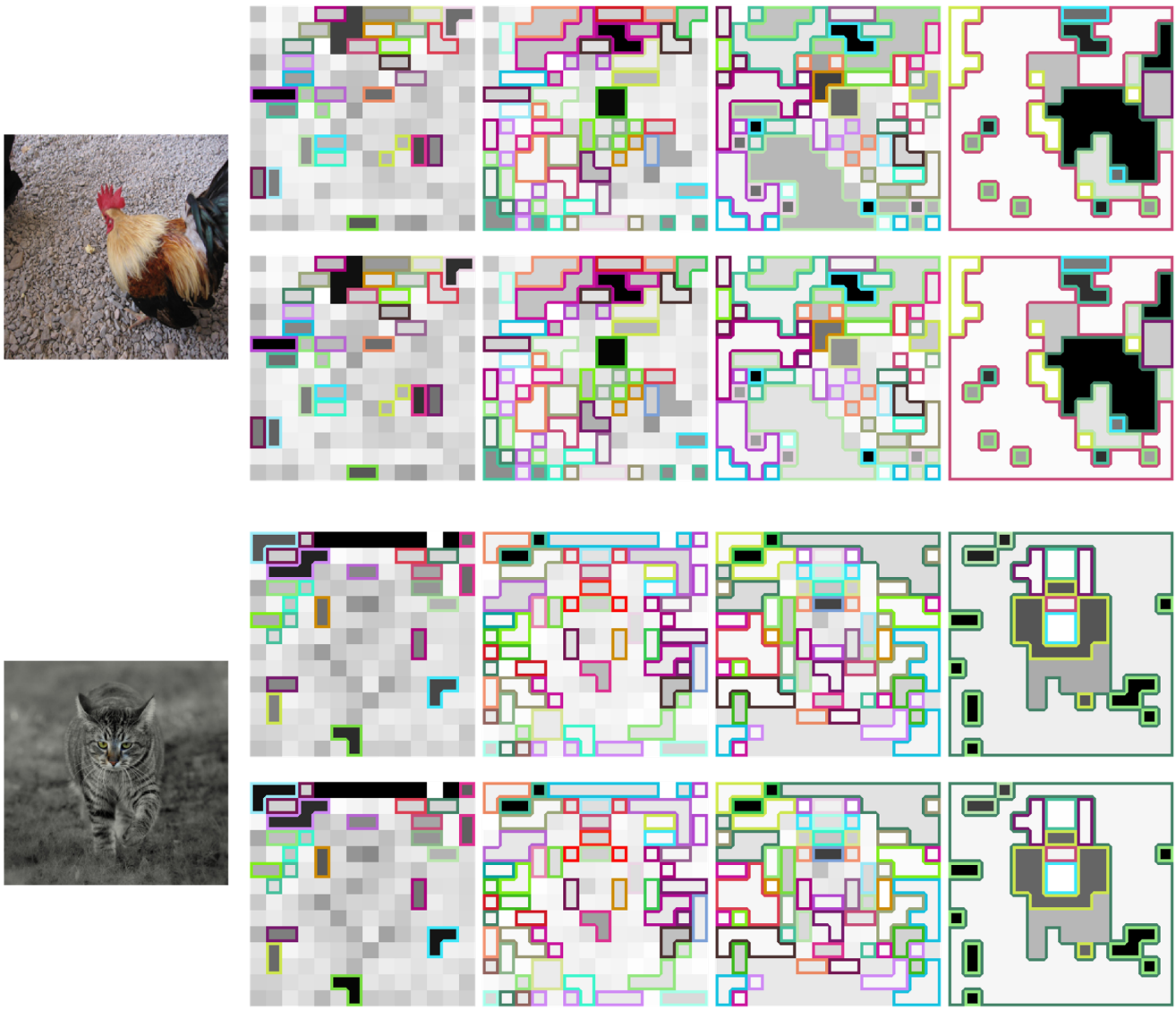

Figure 1. **Comparison of approximated and precise attention map for $\hat{\mathbf{A}}^{l+1}$.** Given the left image, we visualize the (Top) approximated attention map and (Bottom) precise attention map.

C. Detailed results

In this section, we provide more detailed results of MCTF with the Vision Transformers in ImageNet-1K [2].

C.1. Full results with DeiT [7].

As the settings in the ablations studies of the main paper, we first finetune the model with $r = 16$ for the number of reduced tokens per layer and report the flops and accuracies with varying r. We highlight the row used for finetuning. Also, we present the detailed results of MCTF without any additional training. Full results with and without finetuning are summarized in Table C and Table D, respectively.

Table C. **Detailed results of MCTF with DeiT after finetuning with $r = 16$.**

r	DeiT-T				DeiT-S			
	FLOPs		Top-1 Acc		FLOPs		Top-1 Acc	
	(G)	↓ (%)	(%)	Δ	(G)	↓ (%)	(%)	Δ
Base	1.26	-	72.2	-	4.61	-	79.8	-
1	1.24	1.59	72.92	+0.72	4.52	1.95	80.06	+0.26
2	1.20	4.76	72.91	+0.71	4.39	4.77	80.07	+0.27
3	1.17	7.14	72.92	+0.72	4.25	7.81	80.04	+0.24
4	1.13	10.32	72.91	+0.71	4.12	10.63	80.02	+0.22
5	1.09	13.49	72.92	+0.72	3.99	13.45	80.03	+0.23
6	1.06	15.87	72.92	+0.72	3.86	16.27	80.04	+0.24
7	1.02	19.05	72.91	+0.71	3.73	19.09	80.03	+0.23
8	0.98	22.22	72.94	+0.74	3.60	21.91	80.03	+0.23
9	0.95	24.60	72.86	+0.66	3.48	24.51	80.04	+0.24
10	0.91	27.78	72.77	+0.57	3.35	27.33	80.01	+0.21
11	0.88	30.16	72.81	+0.61	3.22	30.15	80.03	+0.23
12	0.84	33.33	72.76	+0.56	3.10	32.75	80.02	+0.22
13	0.81	35.71	72.73	+0.53	2.97	35.57	80.04	+0.24
14	0.78	38.10	72.71	+0.51	2.85	38.18	80.02	+0.22
15	0.74	41.27	72.72	+0.52	2.72	41.00	80.02	+0.22
16	0.71	43.65	72.66	+0.46	2.60	43.60	80.07	+0.27
17	0.68	46.03	72.38	+0.18	2.49	45.99	79.93	+0.13
18	0.65	48.41	72.07	-0.13	2.38	48.37	79.87	+0.07
19	0.62	50.79	71.86	-0.34	2.28	50.54	79.81	+0.01
20	0.60	52.38	71.35	-0.85	2.19	52.49	79.54	-0.26

Table D. **Detailed results of MCTF with DeiT without any addtional training.**

r	DeiT-T				DeiT-S			
	FLOPs		Top-1 Acc		FLOPs		Top-1 Acc	
	(G)	$\downarrow$ (%)	(%)	Δ	(G)	$\downarrow$ (%)	(%)	Δ
Base	1.26	-	72.2	-	4.61	-	79.8	-
1	1.24	1.59	72.15	-0.05	4.52	1.95	79.78	-0.02
2	1.20	4.76	72.09	-0.11	4.39	4.77	79.81	+0.01
3	1.17	7.14	72.06	-0.14	4.25	7.81	79.79	-0.01
4	1.13	10.32	72.06	-0.14	4.12	10.63	79.83	+0.03
5	1.09	13.49	72.06	-0.14	3.99	13.45	79.81	+0.01
6	1.06	15.87	72.00	-0.20	3.86	16.27	79.74	-0.06
7	1.02	19.05	72.00	-0.20	3.73	19.09	79.72	-0.08
8	0.98	22.22	71.98	-0.22	3.60	21.91	79.76	-0.04
9	0.95	24.60	71.92	-0.28	3.48	24.51	79.68	-0.12
10	0.91	27.78	71.88	-0.32	3.35	27.33	79.64	-0.16
11	0.88	30.16	71.82	-0.38	3.22	30.15	79.61	-0.19
12	0.84	33.33	71.72	-0.48	3.10	32.75	79.62	-0.18
13	0.81	35.71	71.61	-0.59	2.97	35.57	79.54	-0.26
14	0.78	38.10	71.50	-0.70	2.85	38.18	79.41	-0.39
15	0.74	41.27	71.28	-0.92	2.72	41.00	79.36	-0.44
16	0.71	43.65	70.99	-1.21	2.60	43.60	79.21	-0.59
17	0.68	46.03	70.62	-1.58	2.49	45.99	79.06	-0.74
18	0.65	48.41	70.01	-2.19	2.38	48.37	78.80	-1.00
19	0.62	50.79	69.41	-2.79	2.28	50.54	78.63	-1.17
20	0.60	52.38	68.52	-3.68	2.19	52.49	78.06	-1.74

C.2. Full results with T2T-ViT [10] and LV-ViT [3].

We also present the full results with T2T-ViT and LV-ViT in Table E. Note that, similar to DeiT-S, we report the FLOPs and accuracies in varying reduction ratios with the model finetuned with a specific reduction ratio, which is used for reporting the results in Table 2 of the main paper. We also highlight this reduction ratio in the table. It is worth noting that, although each model is finetuned with the specific r, MCTF shows promising performance within the range from 1 to r.

Table E. **Detailed results of MCTF with T2T-ViT and LV-ViT.**

r	T2T-ViT$_t$-14				T2T-ViT$_t$-19				LV-ViT-S			
	FLOPs		Top-1 Acc		FLOPs		Top-1 Acc		FLOPs		Top-1 Acc	
	(G)	$\downarrow$ (%)	(%)	Δ	(G)	$\downarrow$ (%)	(%)	Δ	(G)	$\downarrow$ (%)	(%)	Δ
Base	6.11	-	81.7	-	9.81	-	82.4	-	6.50	-	83.3	-
1	6.00	1.80	81.84	+0.14	9.50	3.16	82.42	+0.02	6.34	2.46	83.51	+0.21
2	5.84	4.42	81.85	+0.15	9.10	7.24	82.43	+0.03	6.14	5.54	83.53	+0.23
3	5.69	6.87	81.82	+0.12	8.71	11.21	82.40	$\pm$0.00	5.93	8.87	83.50	+0.20
4	5.53	9.49	81.83	+0.13	8.32	15.19	82.43	+0.03	5.73	11.85	83.51	+0.21
5	5.38	11.95	81.83	+0.13	7.94	19.06	82.39	-0.01	5.52	15.08	83.48	+0.18
6	5.23	14.40	81.84	+0.14	7.56	22.94	82.43	+0.03	5.32	18.15	83.48	+0.18
7	5.07	17.02	81.84	+0.14	7.18	26.81	82.41	+0.01	5.12	21.23	83.52	+0.22
8	4.92	19.48	81.80	+0.10	6.81	30.58	82.42	+0.02	4.93	24.15	83.47	+0.17
9	4.78	21.77	81.81	+0.11	6.44	34.35	82.39	-0.01	4.73	27.23	83.48	+0.18
10	4.63	24.22	81.76	+0.06	6.08	38.02	82.27	-0.13	4.54	30.15	83.47	+0.17
11	4.48	26.68	81.81	+0.11	5.74	41.49	82.25	-0.15	4.35	33.08	83.44	+0.14
12	4.34	28.97	81.80	+0.10	5.45	44.44	82.02	-0.38	4.16	36.00	83.37	+0.07
13	4.19	31.42	81.76	+0.06	5.21	46.89	81.86	-0.54	3.98	38.77	83.23	-0.07
14	4.05	33.72	81.69	-0.01	5.00	49.03	81.38	-1.02	3.83	41.08	83.03	-0.27
15	3.92	35.84	81.51	-0.19	4.82	50.87	80.85	-1.55	3.69	43.23	82.72	-0.58
16	3.80	37.81	81.48	-0.22	4.67	52.40	80.46	-1.94	3.58	44.92	82.28	-1.02
17	3.70	39.44	81.22	-0.48	4.53	53.82	80.29	-2.11	3.48	46.46	81.81	-1.49
18	3.61	40.92	80.93	-0.77	4.41	55.05	79.58	-2.82	3.38	48.00	81.01	-2.29
19	3.53	42.23	80.67	-1.03	4.30	56.17	79.29	-3.11	3.31	49.08	80.73	-2.57
20	3.45	43.54	80.11	-1.59	4.20	57.19	78.41	-3.99	3.23	50.31	79.85	-3.45

References

[1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *ICLR*, 2022. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 4

[3] Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. *NeurIPS*, 2021. 1, 6

[4] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, et al. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *ECCV*, 2022. 1

[5] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *ICLR*, 2017. 1

[6] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *NeurIPS*, 2021. 1

[7] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1, 4

[8] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 2020. 1

[9] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *CVPR*, 2022. 1

[10] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, 2021. 1, 6

[11] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 1

[12] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1