Multi-criteria Token Fusion with One-step-ahead Attention for Efficient Vision Transformers

Sanghyeok Lee* **Joonmyung Choi*** **Hyunwoo J. Kim†**

Department of Computer Science and Engineering, Korea University
{cat0626, pizard, hyunwoojkim}@korea.ac.kr

Abstract

*Vision Transformer (ViT) has emerged as a prominent backbone for computer vision. For more efficient ViTs, recent works lessen the quadratic cost of the self-attention layer by pruning or fusing the redundant tokens. However, these works faced the speed-accuracy trade-off caused by the loss of information. Here, we argue that token fusion needs to consider diverse relations between tokens to minimize information loss. In this paper, we propose a Multi-criteria Token Fusion (MCTF), that gradually fuses the tokens based on multi-criteria (*i.e.*, similarity, informativeness, and size of fused tokens). Further, we utilize the one-step-ahead attention, which is the improved approach to capture the informativeness of the tokens. By training the model equipped with MCTF using a token reduction consistency, we achieve the best speed-accuracy trade-off in the image classification (ImageNet1K). Experimental results prove that MCTF consistently surpasses the previous reduction methods with and without training. Specifically, DeiT-T and DeiT-S with MCTF reduce FLOPs by about 44% while improving the performance (+0.5%, and +0.3%) over the base model, respectively. We also demonstrate the applicability of MCTF in various Vision Transformers (*e.g.*, T2T-ViT, LV-ViT), achieving at least 31% speedup without performance degradation. Code is available at* https://github.com/mlvlab/MCTF.

1. Introduction

Vision Transformer [11] (ViT) has been proposed to tackle the vision tasks with self-attention, originally developed for natural language processing tasks. With the advent of ViT, Transformers are the prevalent architectures for a wide range of vision tasks, *e.g.*, classification [11, 20, 28, 29, 31], object detection [5, 31, 41], segmentation [20, 27, 33], etc. ViTs, built only with self-attention and MLP, provide great flexibility and impressive performance compared to conventional approaches, *e.g.*, convolutional neural networks (CNNs). However, despite these advantages, the quadratic computational complexity of self-attention with respect to the number of tokens is the major bottleneck for Transformers. This limitation becomes more substantial with the growing interest in large-scale foundation models such as CLIP [25]. To this end, several works [3, 16, 30, 34] have proposed efficient self-attention mechanisms including local self-attention within predefined windows [1, 9, 20].

More recently, there has been increasing interest in token-reduction methods for optimizing ViTs without altering their architecture. Earlier works [12, 23, 24, 26, 37] primarily focused on pruning the uninformative tokens to reduce the number of tokens. Another line of works [4, 17, 18, 21, 22] attempted to fuse the tokens instead of discarding them to minimize the information loss. However, performance degradation is still commonly observed in most token fusion methods. We notice that the token fusion methods usually consider only one criterion, such as the similarity or informativeness of tokens, leading to suboptimal token matching. For instance, similarity-based token fusion is prone to combine the foreground tokens, whereas informativeness-based fusion often merges substantially dissimilar tokens, resulting in collapsed representations. Furthermore, if too many tokens are fused into one token, then information loss is inevitable.

To address the problems, we introduce **M**ulti-**C**riteria **T**oken **F**usion (**MCTF**) that optimizes vision transformers by fusing tokens based on multi-criteria. Unlike previous works that consider a single criterion for token fusion, MCTF measures the relationship between the tokens with multi-criteria as follows; (1) similarity to fuse the redundant tokens, (2) informativeness to reduce the uninformative tokens, (3) the size of the tokens to prevent the large-sized tokens that boost the loss of information. Also, to tackle the inconsistency between attention maps of consecutive layers, we adopt *one-step-ahead attention*, which explicitly estimates the informativeness of

*Equal contribution.
†Corresponding author.

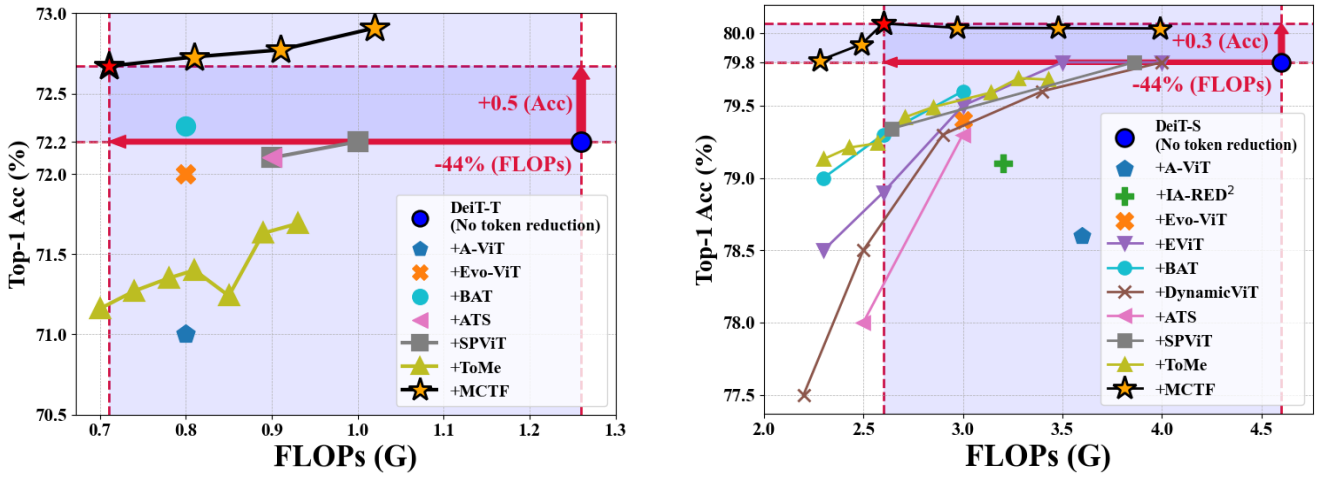


Figure 1. **Comparison of the token reduction methods with DeiT-T (left), and DeiT-S (right).** Given a base model marked as blue circle, previous token reduction methods accelerate the speed with the trade-off between accuracy and computational cost. Our MCTF, marked as a star, even brings performance improvements while lessening the complexity of DeiT. Note that after only one finetuning with the specific reduced number of tokens marked as red star, we simply evaluate it with the diverse FLOPs by adjusting the reduced numbers.

the tokens in the next layer. Finally, by introducing a *token reduction consistency* for finetuning the model, we achieve superior performance to the existing works as in Figure 1. Surprisingly, our MCTF even performs better than the 'full' base model (red dotted line) with a reduced computational complexity. Specifically, it brings a 0.5%, and 0.3% gain while reducing FLOPs by about 44% in DeiT-T, and DeiT-S [28], respectively. We have observed a similar speed-up (31%) in T2T-ViT [39], and LV-ViT [15] without any performance degradation.

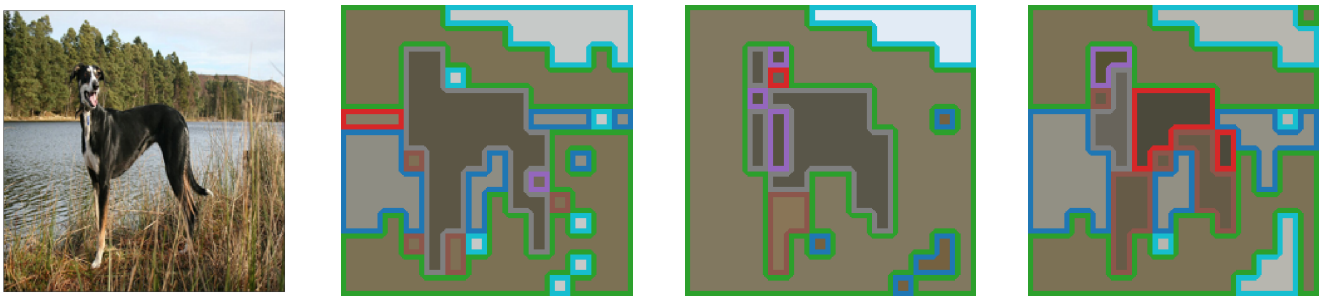
Our contributions are summarized in fourfold.

- We propose *Multi-criteria Token Fusion*, a novel token fusion method that considers multi-criteria, *e.g.*, similarity, informativeness, and size, to capture the complex relationship of tokens and minimize information loss.
- For measuring the informativeness of the tokens, we utilize *one-step-ahead attention* to retain the attentive tokens in the following layers.
- We propose a new fine-tuning scheme with *token reduction consistency* to boost the generalization performance of transformers equipped with MCTF.
- The extensive experiments demonstrate that MCTF achieves the best speed-accuracy trade-off in diverse ViTs, surpassing all previous token reduction methods.

2. Related works

Vision Transformers. Vision Transformer [11] is introduced to tackle the vision tasks. Later, DeiT [28] and CaiT [29] are proposed to handle the data efficiency and scalability of ViT, respectively. Recent works [6, 10, 14, 20, 31] tried to insert the inductive biases of CNNs on ViT, such as the locality or pyramid-architecture. In parallel, there is a line of works that boosts the vanilla ViT by scaling [29, 40] or self-supervised learning [2, 13, 32]. Despite the promising results of these works, the quadratic complexity of ViTs is still the major constraint for scaling the model. For the sake of mitigating the complexity, Reformer [16] lessens the quadratic complexity to $O(N \log N)$ through the hashing function, and Linformer [30], performer [8], and Nyströmformer [34] achieve the linear cost with the approximated linear attention. Also, several works [1, 9, 10, 20] utilize sparse attention with the reduced key or query. Swin [20] and Twins [9] utilize the local attention within the fixed size of the window to mitigate the complexity.

Token reduction in ViTs. Most of the computational burden in ViTs arises from the self-attention. To reduce the quadratic cost in the number of tokens, recent works [4, 12, 17, 18, 21–24, 26, 37] have an interest in reducing the token itself. These works have the advantage of utilizing the original ViTs architecture without modification. In earlier works [12, 23, 24, 26, 37], the uninformative tokens are simply dropped during the forward process, leading to the information loss. To compensate for this, SPViT [17] and EViT [18] first split the tokens into informative and uninformative token sets based on attention scores, then fuse these uninformative token sets into a single token. In parallel, token pooling [22] and ToMe [4] combine the semantically similar tokens to reduce redundancies. A more recent study BAT [21] first split the tokens based on informativeness then fuse the tokens considering the diversity of the tokens. Despite the advantage of each criterion, successful integration of multi-criteria is still less explored.



(a) Origin (b) $\mathbf{W}^{\text{sim}}$ (c) $\mathbf{W}^{\text{sim}}\&\mathbf{W}^{\text{info}}$ (d) $\mathbf{W}^{\text{sim}}\&\mathbf{W}^{\text{info}}\&\mathbf{W}^{\text{size}}$

Figure 2. **Visualization of the fused tokens.** Given (a) the leftmost image, (b) fusing the tokens with a single criterion $\mathbf{W}^{\text{sim}}$ often results in the excessive fusion of the foreground object. (c) Then considering both similarity and informativeness ($\mathbf{W}^{\text{sim}}\&\mathbf{W}^{\text{info}}$), tokens in the foreground objects are less fused while the tokens in the background are largely fused. (d) Finally, MCTF helps retain the information of each component in the image by preventing the large-size token with the multi-criteria ($\mathbf{W}^{\text{sim}}\&\mathbf{W}^{\text{info}}\&\mathbf{W}^{\text{size}}$).

3. Method

We first review the self-attention and token reduction approaches (Section 3.1). Then, we present our multi-criteria token fusion (Section 3.2) that leverages one-step-ahead attention (Section 3.3). Lastly, we introduce a training strategy with token reduction consistency in Section 3.4.

3.1. Preliminaries

In Transformers, tokens $\mathbf{X} \in \mathbb{R}^{N \times C}$ are processed by self-attention defined as

$$\text{SA}(\mathbf{X}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{C}}\right)\mathbf{V}, \tag{1}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{X}\mathbf{W}_{\mathbf{Q}}, \mathbf{X}\mathbf{W}_{\mathbf{K}}, \mathbf{X}\mathbf{W}_{\mathbf{V}}$, and $\mathbf{W}_{\mathbf{Q}}, \mathbf{W}_{\mathbf{K}}, \mathbf{W}_{\mathbf{V}} \in \mathbb{R}^{C \times C}$ are learnable weight matrices. Despite its outstanding expressive power, the self-attention does not scale well with the number of tokens N due to its quadratic time complexity $O(N^2C + NC^2)$. To address this problem, a line of works [12, 23, 24, 26, 37] reduces the number of tokens simply by *pruning* uninformative tokens. These approaches often cause significant performance degradation due to the loss of information. Thus, another line of works [4, 17, 18, 21, 22] *fuses* the uninformative or redundant tokens $\hat{\mathbf{X}} \subset \mathbf{X}$ into a new token $\hat{\mathbf{x}} = \delta(\hat{\mathbf{X}})$, where $\mathbf{X}$ is the set of original tokens, and δ denotes a merging function, *e.g.*, max-pooling or averaging. In this work, we also adopt 'token fusion' rather than 'token pruning' with multiple criteria to minimize the loss of information by token reduction.

3.2. Multi-criteria token fusion

Given a set of input tokens $\mathbf{X} \in \mathbb{R}^{N \times C}$, the goal of MCTF is to fuse the tokens into output tokens $\hat{\mathbf{X}} \in \mathbb{R}^{(N-r) \times C}$, where r is the number of fused tokens. To minimize the information loss, we first evaluate the relations between the tokens based on multi-criteria, then group and merge the tokens through bidirectional bipartite soft matching.

Multi-criteria attraction function. We first define an attraction function $\mathbf{W}$ based on multiple criteria as

$$\mathbf{W}(\mathbf{x}_i, \mathbf{x}_j) = \Pi_{k=1}^{M}(\mathbf{W}^k(\mathbf{x}_i, \mathbf{x}_j))^{\tau_k}, \tag{2}$$

where $\mathbf{W}^k : \mathbb{R}^C \times \mathbb{R}^C \to \mathbb{R}_+$ is an attraction function computed by k-th criterion, and $\tau^k \in \mathbb{R}_+$ is the temperature parameter to adjust the influence of k-th criterion. The higher attraction score between two tokens indicates a higher chance of being fused. In this work, we consider the following three criteria: similarity, informativeness, and size.

Similarity. The first criterion is the similarity of tokens to reduce redundant information. Akin to the previous works [4, 22] requiring the proximity of tokens, we leverage the cosine similarity between the set of tokens for

$$\mathbf{W}^{\text{sim}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2}\left(\frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|} + 1\right). \tag{3}$$

Token fusion with similarity effectively eliminates the redundant tokens, yet it often excessively combines the informative tokens as in Figure 2b, causing the loss of information.

Informativeness. To minimize the information loss, we introduce informativeness to avoid the fusion of informative tokens. To quantify the informativeness, we measure the averaged attention scores $\mathbf{a} \in [0, 1]^N$ in the self-attention layer, which indicates the impact of each token on others: $\mathbf{a}_j = \frac{1}{N}\sum_i^N \mathbf{A}_{ij}$, where $\mathbf{A}_{ij} = \text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_j^T}{\sqrt{C}}\right)$. When $\mathbf{a}_i \to 0$, there's no influence from $\mathbf{x}_i$ to other tokens. With the informativeness scores, we define an informativeness-based attraction function as

$$\mathbf{W}^{\text{info}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\mathbf{a}_i\mathbf{a}_j}, \tag{4}$$

where $\mathbf{a}_i, \mathbf{a}_j$ are the informative scores of $\mathbf{x}_i, \mathbf{x}_j$, respectively. When both tokens are uninformative ($\mathbf{a}_i, \mathbf{a}_j \to 0$),

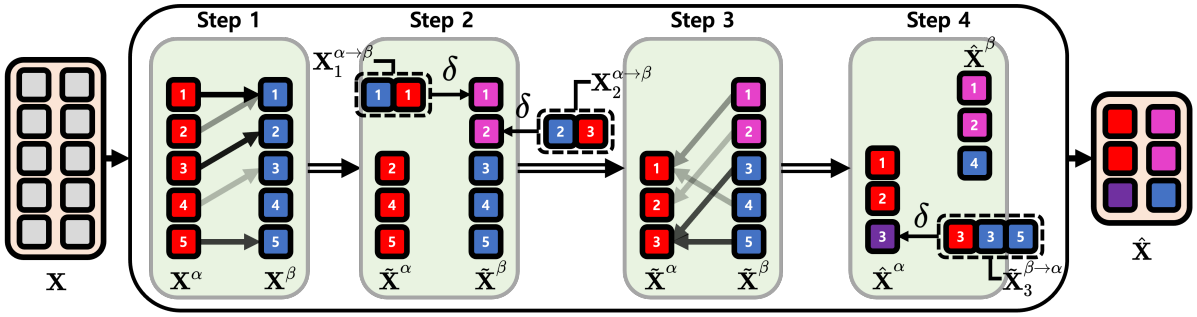


Figure 3. **Bidirectional bipartite soft matching.** The set of tokens $\mathbf{X}$ is split into two groups $\mathbf{X}^\alpha, \mathbf{X}^\beta$, and bidirectional bipartite soft matching are conducted through Step 1-4. The intensity of the lines indicates the multi-criteria weights $\mathbf{W}^t$.

the weight gets higher ($\mathbf{W}^{\text{info}}(\mathbf{x}_i, \mathbf{x}_j) \to \infty$), making two tokens prone to be fused. In Figure 2c, with the weights combined with the similarity and informativeness, the tokens in the foreground object are less fused.

Size. The last criterion is the size of the tokens, which indicates the number of fused tokens. Although tokens are not dropped but merged via a merging function, *e.g.*, averaging pooling or max pooling, it is difficult to preserve all the information as the number of constituent tokens increases. So, the fusion between smaller tokens is preferred. To this end, we initially set the size $\mathbf{s} \in \mathbb{N}^N$ of tokens $\mathbf{X}$ as 1 and track the number of constituent (fused) tokens of each token, and define a size-based attraction function as

$$\mathbf{W}^{\text{size}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\mathbf{s}_i \mathbf{s}_j}. \tag{5}$$

In Figure 2d, tokens are merged based on the multi-criteria: similarity, informativeness, and size. We observed that the fusion happens between similar tokens and the fusion of foreground tokens or large tokens is properly suppressed.

Bidirectional bipartite soft matching. Given the multi-criteria-based attraction function $\mathbf{W}$, our MCTF performs a *relaxed* bidirectional bipartite matching called bipartite soft matching [4]. One advantage of bipartite matching is that it alleviates the quadratic cost of similarity computation between tokens, *i.e.*, $O(N^2) \to O(N'^2)$, where $N' = \lfloor \frac{N}{2} \rfloor$. In addition, by relaxing the one-to-one correspondence constraints, the solution can be obtained by an efficient algorithm. In this relaxed matching problem, the set of tokens $\mathbf{X}$ is first split into the source and target $\mathbf{X}^\alpha, \mathbf{X}^\beta \in \mathbb{R}^{N' \times C}$ as in Step 1 of Figure 3. Given a set of binary decision variables, *i.e.*, the edge matrix $\mathbf{E} \in \{0,1\}^{N' \times N'}$ between $\mathbf{X}^\alpha$, and $\mathbf{X}^\beta$, bipartite soft matching is formulated as

$$\mathbf{E}^* = \arg\max_{\mathbf{E}} \sum\nolimits_{ij} \mathbf{w}'_{ij} \mathbf{e}_{ij} \tag{6}$$

$$\text{subject to} \sum\nolimits_{ij} \mathbf{e}_{ij} = r, \ \sum\nolimits_{j} \mathbf{e}_{ij} \leq 1 \ \forall i, \tag{7}$$

where

$$\mathbf{w}'_{ij} = \begin{cases} \mathbf{w}_{ij} & \text{if } j \neq \arg\max_{j'} \mathbf{w}_{ij'} \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

$\mathbf{e}_{ij}$ indicates the presence of the edge between i, j-th token of $\mathbf{X}^\alpha, \mathbf{X}^\beta$, and , $\mathbf{w}_{ij} = \mathbf{W}(\mathbf{x}_i^\alpha, \mathbf{x}_j^\beta)$. This optimization problem can be solved by two simple steps: 1) find the best edge that maximizes $\mathbf{w}_{ij}$ for each i, and 2) choose the top-r edges with the largest attraction scores. Then, based on the soft matching result $\mathbf{E}^*$, we group the tokens as

$$\mathbf{X}_j^{\alpha \to \beta} = \{\mathbf{x}_i^\alpha \in \mathbf{X}^\alpha \mid \mathbf{e}_{ij} = 1\} \cup \{\mathbf{x}_j^\beta\}, \tag{9}$$

where $\mathbf{X}_i^{\alpha \to \beta}$ indicates the set of tokens matched with $\mathbf{x}_i^\beta$. Finally, the results of the fusion $\tilde{\mathbf{X}}$ are obtained as

$$\tilde{\mathbf{X}} = \tilde{\mathbf{X}}^\alpha \cup \tilde{\mathbf{X}}^\beta, \tag{10}$$

$$\text{where } \tilde{\mathbf{X}}^\alpha = \mathbf{X}^\alpha - \bigcup\nolimits_i^{N'} \mathbf{X}_i^{\alpha \to \beta}, \tag{11}$$

$$\tilde{\mathbf{X}}^\beta = \bigcup\nolimits_i^{N'} \{\delta(\mathbf{X}_i^{\alpha \to \beta})\}, \tag{12}$$

$\delta(\mathbf{X}) = \delta(\{\mathbf{x}_i\}_i) = \sum_i \frac{\mathbf{a}_i \mathbf{s}_i \mathbf{x}_i}{\sum_{i'} \mathbf{a}_{i'} \mathbf{s}_{i'}}$ is the pooling operation considering the attention scores $\mathbf{a}$ and the size $\mathbf{s}$ of the tokens. Still, as shown in Step2 of Figure 3, the number of target tokens $\mathbf{X}^\beta$ cannot be reduced. To handle this issue, MCTF performs *bidirectional bipartite soft matching* by conducting the matching in the opposite direction with the updated token sets $\tilde{\mathbf{X}}^\alpha$, and $\tilde{\mathbf{X}}^\beta$ as in Step 3, 4 of Figure 3. The final output tokens $\hat{\mathbf{X}} = \hat{\mathbf{X}}^\alpha \cup \hat{\mathbf{X}}^\beta$ are defined with the following.

$$\hat{\mathbf{X}}^\alpha = \bigcup\nolimits_i^{N'-r} \{\delta(\tilde{\mathbf{X}}_i^{\beta \to \alpha})\}, \tag{13}$$

$$\hat{\mathbf{X}}^\beta = \tilde{\mathbf{X}}^\beta - \bigcup\nolimits_i^{N'-r} \tilde{\mathbf{X}}_i^{\beta \to \alpha}. \tag{14}$$

Note that calculating the pairwise weights with updated two sets of tokens $\tilde{\mathbf{w}}_{ij} = \mathbf{W}(\tilde{\mathbf{x}}_i^\beta, \tilde{\mathbf{x}}_j^\alpha)$ introduces the additional computational costs of $O(N'(N'-r))$. To avoid this overhead, we approximate the attraction function by the attraction scores before fusion. In short, we just reuse the

Figure 4. **Visualization of attentiveness in consecutive layers.**

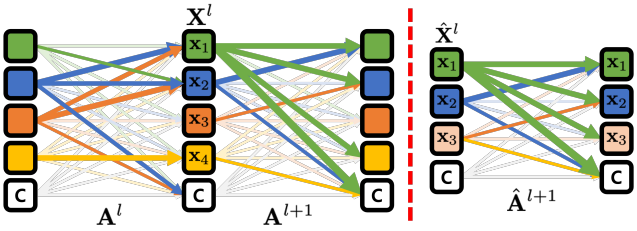


Figure 5. **Illustration of attention map in the consecutive layers and approximated attention.** (Left) The attention score $\mathbf{A}^l$ is the past influence of the tokens to generate $\mathbf{X}^l$. If we fuse the tokens $\mathbf{X}^l$ based on $\mathbf{A}^l$, $\mathbf{x}_1$ is prone to be fused despite the highest informativeness score in the following attention. So, we instead leverage the informativeness based on the one-step-ahead attention $\mathbf{A}^{l+1}$. (Right) After the fusion, we also aggregate the $\mathbf{A}^{l+1}$ to approximate the attention map $\hat{\mathbf{A}}^{l+1}$ for updating fused tokens $\hat{\mathbf{X}}^l$.

pre-calculated weights since $\tilde{\mathbf{X}}^\alpha$ is the subset of $\mathbf{X}^\alpha$. This allows MCTF to efficiently reduce tokens considering bidirectional relations between two subsets with negligible extra costs compared to uni-directional bipartite soft matching.

3.3. One-step-ahead attention for informativeness

In assessing informativeness, prior works [17, 18, 21] have leveraged the attention scores from the previous self-attention layer. As illustrated in Figure 5, previous approaches use the attention $\mathbf{A}^l$ from the previous layer to fuse tokens $\mathbf{X}^l$. This technique allows efficient assessment under the assumption that the attention maps in consecutive layers are similar. However, we observed that the attention maps often substantially differ, as shown in Figure 4, and the attention from a previous layer may lead to suboptimal token fusion. Thus, we proposed **one-step-ahead attention**, which measures the informativeness of tokens based on the attention map in the next layer, *i.e.*, $\mathbf{A}^{l+1}$. Then, the informativeness scores $\mathbf{a}$ in Equation (4) is calculated with $\mathbf{A}^{l+1} \in \mathbb{R}^{N\times N}$. This simple remedy provides a considerable improvement; see Figure 7b in Section 4.2. After token fusion, we efficiently compute the attention map $\hat{\mathbf{A}}^{l+1} \in \mathbb{R}^{(N-r)\times(N-r)}$ of fused tokens $\hat{\mathbf{X}}^l \in \mathbb{R}^{(N-r)\times C}$ by simply aggregating $\mathbf{A}^{l+1} \in \mathbb{R}^{N\times N}$ without recomputing the dot-product self-attention. To be specific, when the tokens are fused as $\delta(\{\mathbf{x}_i\}_i)$ during Equations (10) to (14), their corresponding one-step-ahead attention scores are also fused as $\delta(\{\mathbf{A}_i^{l+1}\}_i)$ in both query and key direction. Note that when fusing attention scores for queries we use simple sum for δ,*i.e.*, $\forall_i \sum_j \hat{\mathbf{A}}_{ij}^{l+1} = 1$. For fusing attention scores for queries, we use simple sum for δ to guarantee $\forall_i \sum_j \hat{\mathbf{A}}_{ij}^{l+1} = 1$.

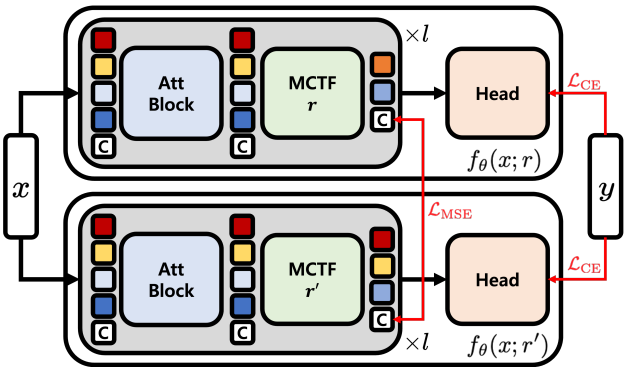


Figure 6. **Illustration of training with token reduction consistency.** During training, we forward the input x as $f(x;r)$, and $f(x;r')$, respectively. To obtain the augmented representation, r' is randomly selected in every step, and the model is updated with supervisory signals $\mathcal{L}_{\text{CE}}$, and consistency loss $\mathcal{L}_{\text{MSE}}$.

3.4. Token reduction consistency

We here propose a new fine-tuning scheme to further improve the performance of vision Transformer $f_\theta(\cdot;r)$ with MCTF. We observe that a different number of reduced tokens per layer, denoted as r, may lead to different representations of samples. By training Transformers with different r and encouraging the consistency between them, namely, token reduction consistency, we achieve the additional performance gain. The objective function of our method is given as

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(f_\theta(x;r), y) + \mathcal{L}_{\text{CE}}(f_\theta(x;r'), y) + \lambda \mathcal{L}_{\text{MSE}}(\mathbf{x}_r^{\text{cls}}, \mathbf{x}_{r'}^{\text{cls}}), \quad (15)$$

where (x, y) is a supervised sample, r, r' is the fixed and dynamic reduced token numbers, λ is the coefficient for consistency loss, and $\mathbf{x}_r^{\text{cls}}, \mathbf{x}_{r'}^{\text{cls}}$ are the class tokens in the last layer of models $f_\theta(x;r), f_\theta(x;r')$. In this objective, we first calculate the cross-entropy loss $\mathcal{L}_{\text{CE}}(f_\theta(x;r), y)$ with fixed r, which is the target reduction number that will be used in the evaluation. At the same time, we generate another representation of the input x with smaller but randomly drawn $r' \sim \text{uniform}(0, r)$, and calculate the loss $\mathcal{L}_{\text{CE}}(f_\theta(x;r'), y)$. Then, we impose the token consistency loss $\mathcal{L}_{\text{MSE}}(\mathbf{x}_r^{\text{cls}}, \mathbf{x}_{r'}^{\text{cls}})$ on the class tokens, to retain the consistent representation across the diverse reduced token numbers r'. The proposed method can be viewed as a new type of token-level data augmentation [7, 19] and consistency regularization. Our token reduction consistency encourages the representation $\mathbf{x}_r^{\text{cls}}$ obtained by the target reduction number r to mimic the slightly augmented representation $\mathbf{x}_{r'}^{\text{cls}}$, which is more similar to ones with no token reduction since $r' < r$.

Table 1. **Image classification results**

Method	FLOPs (G)	Params (M)	Top-1 Acc (%)
DeiT-T [28]	1.2	5	72.2 (-)
+EvoViT$_{[AAAI '22]}$ [36]	0.8	5	72.0 (-0.2)
+A-ViT$_{[CVPR '22]}$ [37]	0.8	5	71.0 (-1.2)
+SPViT$_{[ECCV '22]}$ [17]	0.9	5	72.1 (-0.1)
+ToMe$_{[ICLR '23]}$ [4]	0.7	5	71.3 (-0.9)
+BAT$_{[CVPR '23]}$ [21]	0.8	5	72.3 (+0.1)
+MCTF$_{r=16}$	**0.7**	5	**72.7 (+0.5)**
DeiT-S [28]	4.6	22	79.8 (-)
+IA-RED^2$_{[NeurIPS '21]}$ [24]	3.2	22	79.1 (-0.7)
+DynamicViT$_{[NeurIPS '21]}$ [26]	2.9	23	79.3 (-0.5)
+EvoViT$_{[AAAI '22]}$ [36]	3.0	22	79.4 (-0.4)
+EViT$_{[ICLR '22]}$ [18]	3.0	22	79.5 (-0.3)
+A-ViT$_{[CVPR '22]}$ [37]	3.6	22	78.6 (-1.2)
+ATS$_{[ECCV '22]}$ [12]	2.9	22	79.7 (-0.1)
+SPViT$_{[ECCV '22]}$ [17]	**2.6**	22	79.3 (-0.5)
+ToMe$_{[ICLR '23]}$ [4]	2.7	22	79.4 (-0.4)
+BAT$_{[CVPR '23]}$ [21]	3.0	22	79.6 (-0.2)
+MCTF$_{r=16}$	**2.6**	22	**80.1 (+0.3)**

Table 2. **Comparison with other Vision Transformers**

Models	FLOPs (G)	Params (M)	Acc (%)
PVT-Small[31]	3.8	24.5	79.8
PVT-Medium [31]	6.7	44.2	81.2
CoaT Mini [35]	6.8	10.0	80.8
CoaT-Lite Small [35]	4.0	20.0	81.9
Swin-T [20]	4.5	29.0	81.3
Swin-S [20]	8.7	50.0	83.0
PoolFormer-S36 [38]	5.0	31.0	81.4
PoolFormer-M48 [38]	11.6	73.0	82.5
T2T-ViT$_t$-14 [39]	6.1	21.5	81.7
+MCTF$_{r=13}$	4.2	21.5	**81.8** (↑)
T2T-ViT$_t$-19 [39]	9.8	39.2	82.4
+MCTF$_{r=9}$	6.4	39.2	**82.4** (-)
LV-ViT-S [15]	6.6	26.2	83.3
+EViT$_{[ICLR '22]}$ [18]	4.7	26.2	83.0 (↓)
+BAT$_{[CVPR '23]}$ [21]	4.7	26.2	83.1 (↓)
+DynamicViT$_{[NeurIPS '21]}$ [26]	4.6	26.9	83.0 (↓)
+SPViT$_{[ECCV '22]}$ [17]	4.3	26.2	83.1 (↓)
+MCTF$_{r=12}$	4.2	26.2	**83.4** (↑)

4. Experiments

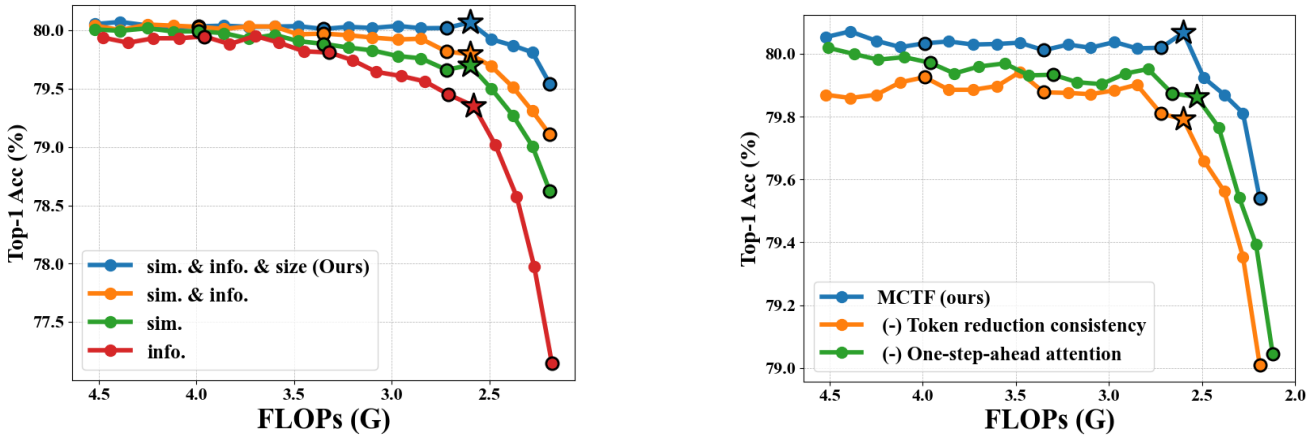
Baselines. To validate the effectiveness of the proposed methods, we compare MCTF with the previous token reduction methods. For comparison, we opt the token pruning methods (A-ViT [37], IA-RED^2 [24], DynamicViT [26], EvoViT [36], ATS [12]) and token fusion methods (SPViT [17], EViT [18], ToMe [4], BAT [21]) in DeiT [28], and report the efficiency (FLOPs (G)) and the performance (Top-1 Acc (%)) of each method. Further, to validate MCTF on other Vision Transformers (T2T-ViT [39], LV-ViT [15]), we report the results of MCTF and compare it with the official number of existing works. We denote the number of reduced tokens per layer r with the subscript in Tables 1 and 2. The gray color in the table indicates the base model, and the green and red color indicates the improvements and degradations of the performance compared to the base model, respectively.

4.1. Experimental Results

Comparison of the token reduction methods. The comparison with existing token reduction methods is summarized in Table 1. We demonstrate that our MCTF achieves the best performance with the lowest FLOPs in DeiT [28] surpassing all previous works. Further, it is worth noting that MCTF is the only work that avoids performance degradation with the lowest FLOPs in both DeiT-T and DeiT-S. Through Finetuning DeiT-T for 30 epochs, MCTF brings a significant gain of +0.5% in accuracy over the base model with nearly half FLOPs. Similarly, we observe a gain of +0.3% with DeiT-S while boosting the FLOPs by -2.0 (G). We believe that multi-criteria with one-step-ahead attention helps the model to minimize the loss of information; further consistency loss on the class token through the token reduction improves the generalizability of the model.

MCTF with other Vision Transformers. To validate the applicability of MCTF in various ViTs, we demonstrate MCTF with other transformer architectures in Table 2. Following previous works [17, 18, 21, 26], we apply MCTF with LV-ViT. Also, we present the results of MCTF with T2T-ViT. As presented in the table, our experimental results are promising. MCTF in these architectures gets at least 31% speedup without performance degradation. Further, MCTF combined with LV-ViT outperforms all other Transformers and token reduction methods regarding FLOPs, and accuracy. Especially, it is worth noting that all token reduction methods except for MCTF bring the performance degradation in LV-ViT. These results reveal that MCTF is the efficient token reduction method for the diverse Vision Transformers.

Token reduction without training. Similar to ToMe [4], MCTF is applicable with pre-trained ViTs without any additional training since MCTF does not require any learnable parameters. We here apply the two reduction methods to the pre-trained DeiT without finetuning and provide the results in Table 3. Regardless of the reduced number of tokens r in each layer, MCTF consistently surpasses ToMe. Especially, in the most sparse setting $r = 20$, the performance gap is significant (+7.0% in DeiT-T, +3.8% in DeiT-S). Note



(a) (b)

Figure 7. **Ablations on (a) multi-criteria, (b) one-step-ahead-attention, and token reduction consistency.** Each marker indicates the model with $r \in [1, 20]$, and we highlight $r \in \{5, 10, 15, 20\}$ as bordered circle. We also denote the model as star when $r = 16$, which is used for finetuning the model.

Table 3. **Image classification results without training**

Method	Base	r = 1	2	4	8	12	16	20
DeiT-T								
ToMe [4]	72.2	72.1	72.0	72.0	71.6	70.8	68.7	61.5
MCTF	72.2	**72.2**	**72.1**	**72.1**	**72.0**	**71.7**	**71.0**	**68.5**
DeiT-S								
ToMe [4]	79.8	79.8	79.7	79.7	79.4	79.0	77.9	74.2
MCTF	79.8	79.8	**79.8**	**79.8**	**79.8**	**79.6**	**79.2**	**78.0**

that without any additional training, our MCTF$_{r=16}$ with pre-trained DeiT-S still shows a competitive performance of 79.2% compared to reduction methods requiring training (*e.g.*, 78.6% of A-ViT, 79.1% of IA-RED^2, and 79.3% of DynamicViT, and SPViT in Table 1).

4.2. Ablation studies on MCTF

We provide ablation studies to validate each component of MCTF. Unless otherwise stated, we conduct whole experiments with DeiT-S finetuned with MCTF ($r = 16$). We provide the FLOPs-Accuracy graph by adjusting the reduced number of tokens per layer $r \in [1, 20]$.

Multi-criteria. We explore the effectiveness of multi-criteria in Figure 7a. First, regarding the multi-criteria, we utilize three criteria for MCTF, *i.e.*, similarity (**sim.**), informativeness (**info.**), and **size**. Each single criterion of similarity and informativeness shows a relatively inferior performance compared to dual (sim. & info.) and multi-criteria (sim. & info. & size). Specifically, when $r = 16$, the performance of a single criterion is 79.7%, and 79.4% with similarity and informativeness, respectively. Then, adopting dual criteria (sim. & info.), MCTF achieves 79.8%. Finally, we get an accuracy of 80.1% with a gain of +0.3% by respecting all three criteria (sim. & info. & size). These performance gaps get larger as r increases, which proves the importance of the multi-criteria for token fusion.

One-step-ahead attention and token reduction consistency. To show the validity of one-step-ahead attention and token reduction consistency, we also provide the results of MCTF with and without each component in Figure 7b. When eliminating either one-step-ahead attention or token reduction consistency, the accuracies are dropped in every FLOP. This significant drop indicates that both approaches matter for MCTF. In short, by adopting one-step-ahead attention and token reduction consistency, MCTF effectively mitigates the performance degradation in a wide range of FLOPs.

Comparison of design choices. The ablations on design choices are presented in Table 4. First, our **bidirectional** bipartite matching, which enables capturing the bidirectional relation in two sets, enhances the accuracy compared to **one-way** bipartite matching. Next, for pooling operation δ, the weighted sum considering the size $\mathbf{s}$ and attentiveness $\mathbf{a}$ is a better choice than others like max-pool or average. Lastly, we compare the results with the precise and approximated attention for $\hat{\mathbf{A}}^l$. For precise attention, we just conduct the similarity calculation for one-step-ahead attention and the attention in the self-attention layer after fusion, separately. Otherwise, we approximate it with one-step-ahead attention as described in Section 3.3. As presented in the table, our approximated attention maintains the performance with the substantial improvement in efficiency (-0.4 (G) FLOPs).

4.3. Analyse of MCTF

Qualitative results. For a better understanding of MCTF, we provide the qualitative results of MCTF in Figure 8. We visualize the fused tokens at the last block of DeiT-S on ImageNet-1K and denote the fused tokens by the same border color. As shown in the figure, since the tokens are merged

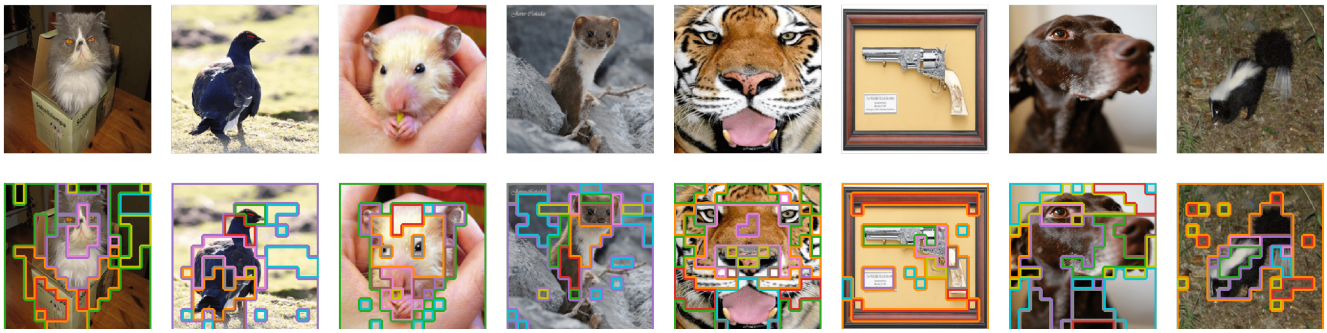


Figure 8. **Visualization of the fused tokens with MCTF.** Given the input images of ImageNet-1K (Top), the qualitative results of MCTF with DeiT-S are provided at the bottom. The same border color of the patches indicates the fused tokens.

Table 4. **Ablations of the design choices.**

Method	FLOPs ↓ (G)	Acc ↑ (%)
DeiT-S	4.6	79.8
bipartite soft matching		
One-way	2.6	80.0
Bidirectional	2.6	80.1
pooling function δ		
average	2.6	80.0
max	2.6	79.8
weighted average	2.6	80.1
approximation of attention map		
precise attention	3.0	80.1
approximated attention	2.6	80.1

with multi-criteria (*e.g.*, similarity, informativeness, size), we maintain the more diverse tokens in the informative foreground object. For instance, in the third image of the hamster, while the background patches including the hand are fused into one token, the foreground tokens are less fused while maintaining the details like the eye, ear, and face of the hamster. In short, compared to the background, the foreground tokens are less fused with the moderate size retaining the information of the main content.

Soundness of size criterion. Figure 9 presents the histogram of sizes of tokens after token reduction with and without size criterion. Specifically, we measure the size of the largest token at the last block and provide the histogram. With our size criterion, the merged tokens tend to have smaller sizes s showing the average size of 39.3/49.2 with and without the Size criterion, respectively. As intended, MCTF successfully suppresses the large-sized tokens, which are a source of information loss, leading to performance improvement.

5. Conclusion

In this work, we introduced the Multi-Criteria Token Fusion (MCTF), a novel strategy aimed at reducing the complexity inherent in ViTs while mitigating performance degradation. MCTF effectively discerns the relation of tokens based on multiple criteria, including similarity, informativeness, and the size of the tokens. Our comprehensive ablation studies and detailed analyses demonstrate the efficacy of MCTF particularly with our innovative one-step-ahead attention and token reduction consistency. Remarkably, DeiT-T and DeiT-S with MCTF achieve considerable improvements, with +0.5%, and +0.3% increase in Top-1 Accuracy over the vanilla models, accompanied by about 44% fewer FLOPs, respectively. We also observe that our MCTF outperforms all of the previous token reduction methods in diverse vision Transformers with and without training.

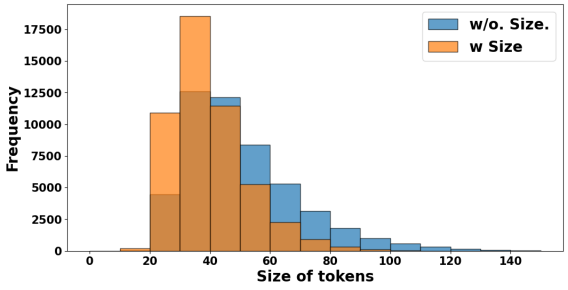


Figure 9. **Histogram of the size of tokens after reduction.**

Acknowledgments

This work was supported by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT)(IITP-2024-2020-0-01819), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(NRF-2023R1A2C2005373), and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare Republic of Korea (HR20C0021).

References

[1] Moab Arar, Ariel Shamir, and Amit H Bermano. Learned queries for efficient local attention. In *CVPR*, 2022. 1, 2

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022. 2

[3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020. 1

[4] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *ICLR*, 2022. 1, 2, 3, 4, 6, 7

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[6] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *ICCV*, 2021. 2

[7] Hyeong Kyu Choi, Joonmyung Choi, and Hyunwoo J. Kim. Tokenmixup: Efficient attention-guided token-level data augmentation for transformers. In *NeurIPS*, 2022. 5

[8] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *ICLR*, 2021. 2

[9] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *NeurIPS*, 2021. 1, 2

[10] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *CVPR*, 2022. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020. 1, 2

[12] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *ECCV*, 2022. 1, 2, 3, 6

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2

[14] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *ICCV*, 2021. 2

[15] 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. 2, 6

[16] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020. 1, 2

[17] 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, 2, 3, 5, 6

[18] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *ICLR*, 2022. 1, 2, 3, 5, 6

[19] Jihao Liu, Boxiao Liu, Hang Zhou, Hongsheng Li, and Yu Liu. Tokenmix: Rethinking image mixing for data augmentation in vision transformers. In *ECCV*, 2023. 5

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1, 2, 6

[21] Sifan Long, Zhen Zhao, Jimin Pi, Shengsheng Wang, and Jingdong Wang. Beyond attentive tokens: Incorporating token importance and diversity for efficient vision transformers. In *CVPR*, 2023. 1, 2, 3, 5, 6

[22] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers for image classification. In *WACV*, 2023. 1, 2, 3

[23] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. Adavit: Adaptive vision transformers for efficient image recognition. In *CVPR*, 2022. 1, 2, 3

[24] Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, and Aude Oliva. IA-RED^2: Interpretability-aware redundancy reduction for vision transformers. *NeurIPS*, 2021. 1, 2, 3, 6

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[26] 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, 2, 3, 6

[27] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 1

[28] 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, 2, 6

[29] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021. 1, 2

[30] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv:2006.04768*, 2020. 1, 2

[31] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense predic-

tion without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 1, 2, 6

[32] QuanLin Wu, Hang Ye, Yuntian Gu, Huishuai Zhang, Liwei Wang, and Di He. Denoising masked autoencoders help robust classification. In *ICLR*, 2023. 2

[33] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 2021. 1

[34] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *AAAI*, 2021. 1, 2

[35] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *ICCV*, 2021. 6

[36] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *AAAI*, 2022. 6

[37] 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, 2, 3, 6

[38] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *CVPR*, 2022. 6

[39] 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. 2, 6

[40] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, 2022. 2

[41] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2021. 1