ImagePiece: Content-aware Re-tokenization for Efficient Image Recognition

Seungdong Yoa[1], Seungjun Lee[1], Hyeseung Cho[1], Bumsoo Kim[2 †], Woohyung Lim[1 †]

[1]LG AI Research
[2]Chung-ang University
{seungdong.yoa, seungjun.lee, hs.cho, w.lim}@lgresearch.ai, bumsoo@cau.ac.kr

Abstract

Vision Transformers (ViTs) have achieved remarkable success in various computer vision tasks. However, ViTs have a huge computational cost due to their inherent reliance on multi-head self-attention (MHSA), prompting efforts to accelerate ViTs for practical applications. To this end, recent works aim to reduce the number of tokens, mainly focusing on how to effectively prune or merge them. Nevertheless, since ViT tokens are generated from non-overlapping grid patches, they usually do not convey sufficient semantics, making it incompatible with efficient ViTs. To address this, we propose ImagePiece, a novel re-tokenization strategy for Vision Transformers. Following the MaxMatch strategy of NLP tokenization, ImagePiece groups semantically insufficient yet locally coherent tokens until they convey meaning. This simple retokenization is highly compatible with previous token reduction methods, being able to drastically narrow down relevant tokens, enhancing the inference speed of DeiT-S by 54% (nearly 1.5× faster) while achieving a 0.39% improvement in ImageNet classification accuracy. For hyperspeed inference scenarios (with 251% acceleration), our approach surpasses other baselines by an accuracy over 8%.

Introduction

"All that is gold does not glitter, not all those who wander are lost." — J.R.R. Tolkien

Tokenization involves breaking down text into smaller units called tokens, such as words, and is a crucial preprocessing step for nearly all Natural Language Processing (NLP) tasks. Modern NLP models like BERT (Devlin et al. 2018), GPT (Brown et al. 2020), and XLNet (Yang et al. 2019) tokenize text into subword units. These subword units strike a balance between words and characters, preserving linguistic meaning while minimizing out-of-vocabulary issues even with a relatively small vocabulary. Subword tokenizer such as wordpiece (Devlin et al. 2018) employs a greedy longest-match-first strategy, iteratively selecting the longest prefix of the remaining text that matches a vocabulary token, a method known as Maximum Matching or MaxMatch.

†Corresponding authors

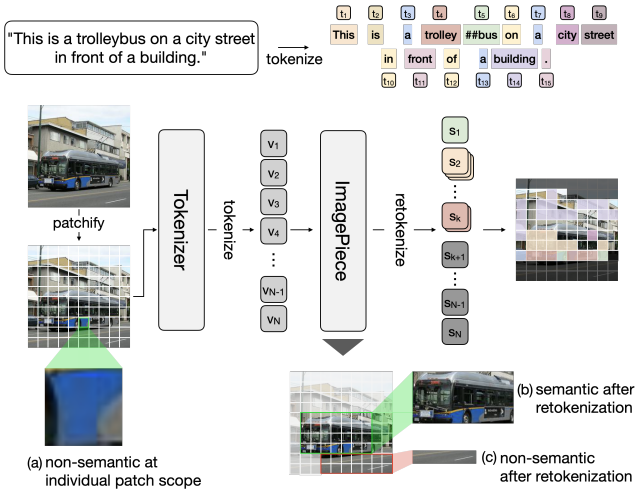

Figure 1: An illustration of the ImagePiece pipeline compared to WordPiece in NLP. While text is tokenized into meaningful tokens, image tokens from patches often contain irrelevant or non-semantic information. The retokenization in ImagePiece enables these non-semantic tokens to be merged into meaningful tokens, particularly when they have the potential to become meaningful after retokenization.

Following the pioneering success of Transformers in natural language processing, Vision Transformers (Yin et al. 2022) have showcased remarkable performance across a broad spectrum of computer vision tasks. In the context of images, tokenization involves splitting an image into non-overlapping grid patches and feeding the sequence of linear embeddings of these patches into a Transformer. Here, image patches are treated similarly to tokens (words) in NLP. However, unlike textual tokens stemmed from subwords where most elements contain meaning that contributes to understanding, image patch-based tokens differ significantly as they often 1) contain totally irrelevant semantics with the overall meaning (e.g., roads, sky, background) or 2) require a broader context to convey meaning (for example, in Fig 1, the blue patch itself (a) hardly contains any semantics while its collection with nearby patches (b) can be correctly recognized as a 'bus').

This phenomenon becomes a more serious issue in further works streamlining ViTs; i.e., *reducing* the number of input tokens (Liu et al. 2021; Dong et al. 2022; Fan et al. 2021; Li et al. 2022; Xiao et al. 2021; Chu et al. 2021; Wang et al. 2021; Graham et al. 2021). In ViTs, token reduction is mainly achieved through two primary methods: 1) *token pruning*: removing inattentive tokens or 2) *token merging*: fusing redundant (or similar) tokens into new abstractions. Both approaches effectively reduce the complexity of Vision Transformers (ViTs). However, due to the limitations of patch-based tokenization, this results in either 1) discarding tokens too hastily before their meaning has fully contextualized, or 2) prematurely smoothing semantic tokens with the nearest-similarity non-semantic noise, rather leading to more serious degradation than token pruning in certain scenes.

To address this, we propose **ImagePiece**, a novel re-tokenization strategy for Vision Transformers. Following the MaxMatch strategy of NLP tokenization (Devlin et al. 2018), ImagePiece groups semantically insufficient (bottom-k) tokens to a level that convey meaning. To add local inductive bias for visual scenes, we add a *local coherence bias* module consisting of overlapping convolution layers to enhance non-semantic patches with similar positions to have higher similarity. This facilitates the re-tokenization to merge nearby non-semantic tokens until they form meaningful tokens (or end up to be irrelevant with the overall meaning). Then, the attention score of the semantic abstraction of inattentive tokens are measured again: if the abstraction gains relevance with the overall visual semantic, the token is re-organized. For further efficiency, tokens that do not participate in the final semantic even after our re-tokenization are discarded.

ImagePiece poses several major advantages: 1) The locally coherent chunks of scattered inattentive tokens are facilitated to be merged to form a semantically meaningful abstraction, enabling the model to perform well-informed contextualization or pruning. 2) The attentive tokens that contain semantically important information are preserved from being merged (or smoothed), mitigating the limitation of previous token-merging pipelines. Benefiting from these advantages, ImagePiece is able to drastically narrow down relevant tokens, enhancing the inference speed of DeiT-S by 54% (nearly 1.5$\times$ faster) while achieving a 0.39% improvement in ImageNet classification accuracy. This is notable given that other baselines in efficient ViTs (Liang et al. 2022; Bolya et al. 2023) suffer from an unignorable degradation in performance as a trade-off for this speed gain. Our contributions are threefold:

- We introduce ImagePiece, a novel retokenization strategy for efficient vision transformers that condense semantically irrelevant tokens. After condensing, the semantics of the tokens are re-evaluated, reorganizing tokens that have obtained meaning by gaining more contextual information.
- With local cohesive bias, non-semantic tokens that are positionally nearby are facilitated to be condensed together, adding local inductive bias of visual scenes. The condensed tokens that have no relevance after several iterations are discarded for efficiency.
- Our method achieves superior performance in ImageNet classification and demonstrates robust performance in hyper-speed inference, outperforming existing baselines.

Related Work

Vision Transformers. Transformers (Vaswani et al. 2017), originally prevalent in NLP, have recently attracted significant attention in computer vision, primarily due to their exceptional ability to model long-range dependencies. Vision Transformers (ViTs) (Dosovitskiy et al. 2021) first introduced Transformer backbones to the field of computer vision, and a diverse array of studies (Touvron et al. 2021a; Yuan et al. 2021; Zhou et al. 2021; Touvron et al. 2021b; Pan et al. 2022; Pan, Cai, and Zhuang 2022; Liu et al. 2021; Dong et al. 2022; Li et al. 2022; Xiao et al. 2021; Wang et al. 2021; Graham et al. 2021) on ViTs have demonstrated their success in various aspects, ranging from architectural improvements to optimization techniques.

One of the core differences between NLP and computer vision applications of transformers lies in how input data is tokenized. In NLP, methods like WordPiece (Devlin et al. 2018) and SentencePiece (Kudo and Richardson 2018) tokenize text into subwords or characters, where each token carries meaningful semantic content. In contrast, when images are split into patches for ViTs, these patches are treated as tokens, yet each individual patch often lacks inherent semantic meaning. This fundamental difference presents unique challenges and opportunities in how transformers are applied across these domains.

Efficient Transformers. Recent advancements in both NLP and computer vision domains have experienced a surge in efforts to enhance transformer models' efficiency. These efforts include developing efficient attention mechanisms (Kitaev, Kaiser, and Levskaya 2020; Bolya et al. 2022; Wang et al. 2020; Dao et al. 2022; Shen et al. 2021; Choromanski et al. 2020), pruning of transformer heads or features (Michel, Levy, and Neubig 2019; Voita et al. 2019; Meng et al. 2022), and integrating vision-specific modules (Liu et al. 2021; Dong et al. 2022; Graham et al. 2021; Mehta and Rastegari 2022). Several recent approaches in NLP (Lassance et al. 2021; Kim et al. 2022; Kim and Cho 2020; Goyal et al. 2020) and computer vision (Meng et al. 2022; Ryoo et al. 2021; Yin et al. 2022; Kong et al. 2022; Song et al. 2022; Fayyaz et al. 2022; Yu and Wu 2023; Marin et al. 2021; Xu et al. 2022; Pan et al. 2021; Long et al. 2023; Liang et al. 2022; Rao et al. 2021; Bolya et al. 2023; Chen et al. 2023) have attempted to reduce the number of tokens due to the input-agnostic nature of transformers. Especially, for ViTs, recent works to reduce the number of tokens have emerged by diverse approaches, such as token pruning and token merging. In token pruning, DynamicViT (Rao et al. 2021) introduces a method to prune tokens for a fully pre-trained ViT using additional training parameters. EViT (Liang et al. 2022) determines inattentive tokens according to class token attention, and discards

these tokens to reorganize image tokens. Token merging approaches, such as ToMe (Bolya et al. 2023), combine the similar token pairs into new tokens to reduce the number of tokens. Our approach provides a novel perspective for efficient ViTs by proposing the retokenization to initially merge non-semantic tokens into semantically meaningful chunks to correctly measure their relevance with the visual scenes.

Preliminary

In this preliminary section, we start with a basic overview of Vision Transformers (ViTs), i.e., how images are patchified and tokenized for Transformer architectures. Next, we explain the basic concepts involved in evaluating each patch token. Finally, we describe the challenges faced by previous token reduction methods.

ViT Overview. ViT first divides an input image into *non-overlapping* $p \times p$ patches and projects each patch to a token embedding. Typically, with a patch size of 16×16 (p=16) and an image size of 224×224, we obtain 196 image tokens. An extra class token, denoted as `[CLS]`, is appended to the sequence of image tokens to serve as an aggregator of global image information for the final classification. After all of the tokens are combined with positional embeddings, the patch embeddings are fed into a transformer encoder.

Evaluating Token Importance. Following previous work in literature (Liang et al. 2022), we define the class attention score as the measure of interaction between the class token and the image tokens, indicating the importance of each token in contributing to the overall semantics of an image. Let D, N denote the length of the query vector and the number of input tokens, respectively, where the input tokens in each transformer block refer to the class token and the remaining image tokens. The token sequence X is projected into a query matrix $\mathbf{Q} \in \mathbb{R}^{N \times D}$, a key matrix $\mathbf{K} \in \mathbb{R}^{N \times D}$, and a value matrix $\mathbf{V} \in \mathbb{R}^{N \times D}$. The class attention score and the output of the class token are as follows:

$$x_{\text{class}} = \text{Softmax}\left(\frac{\mathbf{Q}_{\text{class}} \cdot \mathbf{K}^{\top}}{\sqrt{D}}\right)\mathbf{V} = A_{\text{class}} \cdot \mathbf{V}, \qquad (1)$$

where $\mathbf{Q}_{\text{class}}$ denotes the query vector of the class token. As a result, the output of the class token, denoted as x_{class}, is a linear combination of the value vectors $\mathbf{V}$=$[v_1, v_2, \cdots, v_N]^{\top}$. The coefficients of this combination, denoted by A_{class} in Eq.(1), are the attention values from the class token with respect to all tokens.

Challenges in Token Reduction. Token reduction techniques, including token pruning and token merging, are essential for accelerating ViTs by decreasing the number of tokens. Token pruning (Liang et al. 2022; Rao et al. 2021; Kong et al. 2022; Yin et al. 2022; Meng et al. 2022) and token merging (Bolya et al. 2023; Ryoo et al. 2021; Marin et al. 2021) are the two most representative branches of ViT acceleration. Token pruning methods eliminate less attentive tokens based on a fixed hyper-parameter, targeting the fixed bottom-k tokens deemed unimportant or those under a predetermined threshold, as discussed in (Liang et al. 2022; Meng et al. 2022). However, since tokens are either treated as important or non-important according to the attentiveness score with the `[CLS]` token, locally coherent tokens that might entail semantically crucial information are often discarded. To compensate the information loss that often occurs in token pruning, token merging approaches attempted to accelerate ViTs by merging pairs of tokens with the highest similarity, selecting the top-k most similar pairs rather than removing original tokens. However, as the merging process progresses, tokens with low similarity are prone to be merged, resulting in semantically essential tokens to be diluted to a semantically irrelevant interpolation of two unrelated representations.

ImagePiece

We introduce ImagePiece, a novel re-tokenization strategy that makes image patch tokens analogous to subwords (minimal unit that conveys meaning). Fig. 2 illustrates the overall architecture of our proposed ImagePiece.

Re-tokenizing Non-semantic Tokens. Following the MaxMatch strategy of WordPiece tokenization (Devlin et al. 2018), ImagePiece performs retokenization for ViTs that groups semantically insufficient (bottom-k) tokens to a level that conveys meaning. In detail, the re-tokenization iteratively follows a three-stage procedure. Step I, the importance of each token to the overall semantic is evaluated by the attention score with the `[CLS]` token (Liang et al. 2022) as in Eq.(1). Step II, the bottom-k tokens are divided into two equal-sized subsets, where the tokens organized by their attention are alternatively assigned to each group (for convenience, we refer to these groups as A and B). Then, each token in group A is merged with the most similar token in group B using bipartite soft matching (Bolya et al. 2023). Step III, the attention scores of the semantic abstractions of the matched tokens along with remaining non-bottom-k tokens are recalculated: the abstraction tokens that gain relevance with the overall visual semantic are re-organized.

Local Coherence Bias. As illustrated in Fig. 1, patches whose meaning is hard to discriminate at an individual scope can gather meaningful semantics when contextualized with nearby patches (i.e., local inductive bias). To add this bias to visual tokens, we add a *local coherence bias* module consisting of overlapping convolution layers. As adjacent patches are entangled with overlapping features, non-semantic patches that are geometrically close are encouraged to have higher similarity. This facilitates the retokenization to merge nearby non-semantic tokens until they form meaningful tokens (or end up being irrelevant with the overall meaning).

Compatibility with Token-Pruning. As ViTs suffer from the quadratic complexity of input tokens and the square-grid patch tokenizer (Yin et al. 2022) usually contains high portions of irrelevant or redundant patches, previous works have devised pruning strategies to *drop-out* tokens that do not contribute to the overall semantic of a visual scene (Rao et al. 2021; Liang et al. 2022). However, patches whose semantic is difficult to determine without proper contextual-

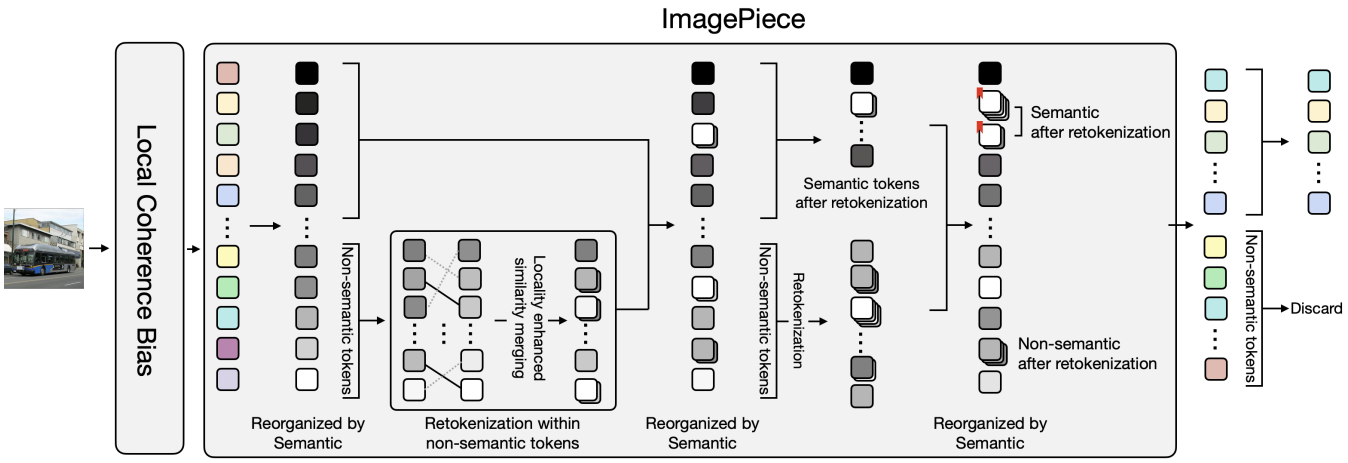

Figure 2: Overall architecture of the proposed method.

ization are prone to be discarded at earlier layers due to the progressively pruning strategy where the tokens are discarded at intermediate ViT layers. On the other hand, due to this progressive architecture, ImagePiece shows powerful compatibility with existing token pruning methods, since the non-semantic tokens are quickly merged with geometrically nearby tokens, benefiting from strong local bias to form meanings at earlier transformer layers (see Fig. 3).

Competitiveness with Token-Merging. ImagePiece can be treated as a variation of token merging (Bolya et al. 2023; Ryoo et al. 2021; Marin et al. 2021). Previous work in token merging is mainly focused on combining redundant tokens (with high feature similarities) to enhance efficiency in Vision Transformers. However, this strategy entails serious drawbacks. For instance, when the similarity between the top-k similar token pairs are not sufficient enough, the merging pipelines above results in *smoothing* highly semantic tokens with rather irrelevant tokens. This leads to rather harming crucial semantics of the visual scene. Tab.7 shows empirical support that semantically-relevant tokens are frequently merged (thus being semantically diluted) in the initial ViT layers with previous token merging methods. Conversely, ImagePiece retokenizes non-semantic tokens until they contribute to a meaningful semantic representation. Experimental results show that due to this strategical difference, ImagePiece successfully preserves semantically essential tokens without interfering with highly-semantic tokens.

Experiments

In this section, we conduct extensive experiments to demonstrate our retokenization strategy for efficient ViTs. We first show that our method outperforms the baselines by a substantial margin. Analytical experiments are conducted to further explain the factors contributing to our method's effectiveness. Notably, the hyper-speed inference experiment in Fig. 3 reveals that our method achieves relatively robust performance even with a drastically reduced number of tokens.

Implementation details. We conduct all the experiments on ImageNet-1k (Deng et al. 2009) consisting of 1.2 million images in the training set and 50k images in the test set. The image resolution is 224×224 in training and testing. During training our models, we simply follow all the training strategies and optimization methods used in DeiT (Touvron et al. 2021a). We train our model from scratch for 300 epochs, and we don't use any tricks (e.g., adding extra parameters, starting from an existing checkpoint or fine-tuning, using additional training tricks), unlike other prior works. The throughput (img/s) is measured on a single NVIDIA GeForce RTX 3090 during inference.

For our method to apply the local coherence bias module, we adopt simple convolutional layers (four 3×3 convolutions and a single 1×1 convolution), replacing a standard ViT's patchify stem. In all experiments, we set the proportion p of the non-semantic token set to 0.3, targeting the bottom 30% of tokens based on their importance (attentiveness). Therefore, these tokens are candidates for retokenization. We set the similarity merging ratio to 0.08, meaning that token pairs equivalent to 0.08 $\times$ the total number of tokens from the non-semantic token set are selected for merging based on their similarity. Also, we set the pruning ratio r to 0.8, which results in discarding the bottom 20% of tokens, identified as non-semantic, after retokenization by ImagePiece. These hyperparameters are specifically adjusted to further accelerate ViTs beyond the standard settings.

Main Results

In Tab. 1 and Tab. 2, we describe the main results on ImageNet (Deng et al. 2009) using the base models, DeiT-Ti and DeiT-S. We report the top-1 accuracy (%) and throughput (img/s).

Comparisons with prior pruning methods. To demonstrate the effectiveness of our retokenization approach in pruning, we compare our ImagePiece with prior pruning methods. For token pruning, DynamicViT (Rao et al. 2021)

Model	Acc (%)	Throughput (img/s)
DeiT-Ti	72.13	6429.2
Pruning by learned projection layer (Rao et al. 2021)	71.20 (-0.93)	9351.0
Pruning by learned token selector (Kong et al. 2022)	72.09 (-0.04)	9245.2
Pruning by [CLS] attentiveness (Liang et al. 2022)	71.70 (-0.43)	9432.9
Pruning by retokenization (Ours)	**72.61** (+0.48)	**9450.2**
DeiT-S	79.83	2531.1
Pruning by learned projection layer (Rao et al. 2021)	79.32 (-0.51)	3762.0
Pruning by learned token selector (Kong et al. 2022)	79.34 (-0.39)	3722.1
Pruning by [CLS] attentiveness (Liang et al. 2022)	79.37 (-0.46)	3787.5
Pruning by retokenization (Ours)	**80.22** (+0.39)	**3891.9**

Table 1: Comparison to existing token pruning models with DeiT on ImageNet-1k. The color of the number indicates the performance gap compared to the original models, DeiT-Ti and DeiT-S. Ours outperforms all the baselines, especially the original DeiT-S improving throughput (img/s) by 54%.

employs a *learned projection layer*, SPViT (Kong et al. 2022) utilizes a *learned token selector*, and EViT (Liang et al. 2022) leverages a *[CLS] attentiveness score*. Overall, our method achieves the best accuracy with slightly faster inference speed as shown in Tab. 1. Especially, compared to the original DeiT-Ti and DeiT-S, our method accelerates the inference speed by 47% and 54% respectively, while also notably improving the performance with an accuracy gap of 0.48% for DeiT-Ti and 0.39% for DeiT-S.

Comparisons to token merging models. As shown in Tab. 7, our observations reveal that important tokens are frequently merged in the initial layers of previous token merging method. This indicates that important tokens are quickly merged in the early layers, and as the merging process progresses to later layers, the proportion of inattentive tokens increases, leading to lower similarity in the merging process at these stages. Additionally, since merging tokens can potentially condense the information of target tokens, it may also reduce the granularity of details, particularly in important tokens, leading to a loss of critical information. To address these issues, our method, ImagePiece, selectively merges only the less significant tokens.

In Tab. 2, we compare our method with existing token merging methods: ToMe (Bolya et al. 2023) which uses bipartite soft matching algorithm for merging, Token Pooling (Marin et al. 2021) which is similar to ToMe (Bolya et al. 2023) but utilizes a slow K-Means based approach, and Token Learner (Ryoo et al. 2021) which uses an MLP to reduce the number of tokens. Unlike these methods, our approach focuses on transforming non-semantic tokens into semantically meaningful ones. This crucial step in our retokenization process, implemented through ImagePiece, provides a significant advantage over conventional token merging strategies. Tab. 2 shows that even without pruning, ImagePiece outperforms these baseline methods with faster inference speed. Our method, which employs a pruning process after the retokenization of ImagePiece, achieves the best accuracy, outperforming existing token merging methods by 3.65% on average.

Model	Acc (%)	Throughput (img/s)
Token Pooling (Marin et al. 2021)	71.35 (-8.48)	3571.0
Token Learner (Ryoo et al. 2021)	79.01 (-0.82)	3747.7
ToMe (Bolya et al. 2023)	79.36 (-0.47)	3806.1
ImagePiece	79.67 (-0.16)	3890.1
ImagePiece + Pruning (Ours)	**80.22** (+0.39)	**3891.9**

Table 2: Comparison to prior token merging methods with DeiT-S on ImageNet-1k. Ours outperforms all the baseline models by an average of 3.65%.

Extensive Experimental Results

We also conduct extensive experiments, such as hyper-speed evaluation and random masking noise evaluation, to demonstrate the efficiency and effectiveness of our approach. All experiments are conducted on 50k images from the ImageNet test set, and unless otherwise stated, the inference speed for all models is maintained at the same level.

Hyper-speed Inference Results. We conduct a hyper-speed evaluation experiment to demonstrate the effectiveness of our method in Fig. 3. In this experiment, we use the models trained in typical settings mentioned in Tab. 1 and Tab. 2, e.g., DynamicViT with a keep rate of 0.7, EViT with a keep rate of 0.7, ToMe with a reduction factor r=13 (the number of merging tokens per layer), and ours with a keep rate of 0.8 and a merging rate of 0.08. Using these models, we evaluate their performance at various higher speeds without further training. For example, EViT initially trained at a keep rate of 0.7 is assessed at lower keep rates such as 0.6, 0.5, . . . , 0.27, and ToMe initially trained at r=13 is evaluated at higher r values (e.g., 16, 18, . . . , 25) to further accelerate the inference speed. The details of each model's setting are in the supplements.

The performance gap between the baselines and ours is gradually increasing (0.86% to 8.15% on average) as tokens are further reduced to accelerate the inference speed. Specifically, when compared to DeiT-S, at a 2.51× speedup, the accuracy of DynamicViT, EViT, and ToMe decreases by 20.05%, 13.99%, and 9.24% respectively, whereas our method shows only a 6.28% decrease. This highlights the superior efficiency of our approach in relatively maintaining accuracy under significant token reduction. Additionally, Tab. 3 demonstrates that our method achieves 79.38% accuracy with only 26 output tokens for prediction, which represents just 13% of the original total of 197 tokens (including `[CLS]`), compared to 68 (35%), 69 (35%), and 41 (21%) output tokens for DynamicViT, EViT, and ToMe respectively. This indicates that at a similar performance level, our method improves speed by approximately 30% over the baselines with fewer tokens. From the experiments, we demonstrate that our method successfully retokenizes semantically insufficient tokens into more meaningful units and effectively discards non-semantic tokens by more accurate re-evaluation. This enables it to maintain relatively robust performance, even under drastic token reduction for faster inference speed.

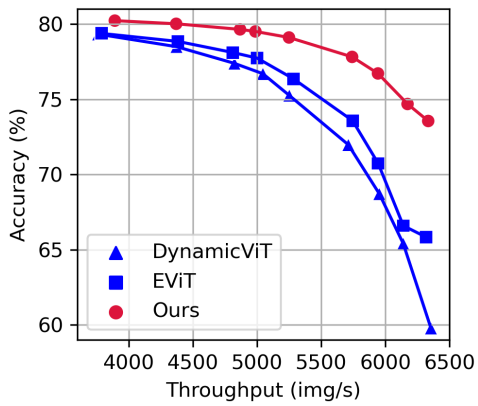

Figure 3: Comparison of our ImagePiece with the patch tokenizer of ViT in terms of inference speed and accuracy. While the baselines show a marked decline in performance as throughput increases, our method not only maintains relatively robust accuracy but also demonstrates compatibility, even at high speeds.

Model Robustness to Random Masking Noise. We also conduct a random masking noise experiment to further investigate the effectiveness of our retokenization method. In Tab. 4, we report the performance of each method on randomly masked test samples. 16×16 masks are applied to the test samples at random locations, and the evaluation is performed as the number of masks gradually increases. In this experiment, we use the models in Tab. 1 and Tab. 2, based on DeiT-S, conducting evaluations through inference.
When the number of random masks is 7, only our method maintains performance over 79.5%, with a performance gap of 1.4% between ours and the baselines. As the number of masks increases, this gap also increases; notably, at 50 masks, the performance gap between our method and the baselines averages 5.2%. This suggests that semantically meaningful tokens by retokenization are more helpful in comprehending a global visual scene with masking noise than individual patch-scope tokens, as it can achieve better performance than the baselines even with just a few non-masked areas of the image. Additionally, while the baselines exhibit a performance of around 79.3% without noise in Tab. 1 and Tab. 2, our method surpasses this, reaching 79.63%, even with 7 random masks present.

Analysis

Token Attentiveness after Re-tokenization. We propose ImagePiece, which is a novel retokenization strategy for token reduction. ImagePiece initially merges non-semantic tokens into semantically meaningful chunks to accurately assess their relevance within the global scene of the image. After merging the tokens from the inattentive token set in a previous layer, we then discard the tokens considered inattentive by evaluating the attentiveness of all tokens. Although some tokens are identified as inattentive in the previous layer, they can be reassessed as attentive tokens after the retokenization if they embody more semantic information as chunks. Tab. 5 demonstrates how many tokens initially treated as inattentive can be reassessed as attentive when merged into chunks. The ratio $\text{Ratio}_{\text{inattn}\rightarrow\text{attn}}$ in Tab. 5 represents the proportion of tokens that were initially deemed inattentive and subsequently retokenized in the previous layer, but are now evaluated as attentive in the current layer. For example, 27.99% of tokens deemed inattentive and retokenized in the 1-st layer are reassessed as attentive in the 2-nd layer, highlighting the capability of ImagePiece to reevaluate token significance. This reevaluation is particularly pronounced in the 6-th layer, where 58.77% of tokens that were considered inattentive and merged in the 5-th layer are reassessed as attentive.

Model	Acc	Throughput	#output tokens (% of Total)
DynamicViT	79.32	3762.0	68 (35%)
EViT	79.37	3787.5	69 (35%)
ToMe	79.36	3806.1	41 (21%)
Ours	**79.38**	**4868.0**	**26** (13%)

Table 3: Comparison of throughput (img/s) and the number of output tokens after token reduction at a similar performance level. Our method achieves approximately 30% speed increase while retaining accuracy comparable to the baselines, with only 26 output tokens remaining, representing 13% of the original total tokens.

Model	# of Random Masks					
	7	10	15	20	25	50
DeiT-S	79.05	78.65	77.78	76.52	75.29	68.28
DynamicViT	78.54	78.06	77.09	75.95	74.81	68.17
EViT	78.21	77.72	76.77	75.42	74.47	68.10
ToMe	78.41	77.73	76.49	74.99	73.68	65.41
Ours	**79.63**	**79.13**	**78.36**	**77.20**	**76.63**	**71.42**

Table 4: Random masking noise evaluation: comparison to baseline models on randomly masked test samples.

Layer	2	3	5	6	8	9
$\text{Ratio}_{\text{inattn}\rightarrow\text{attn}}$ (%)	27.99	44.39	35.49	58.77	28.97	43.27

Table 5: The percentage of previously inattentive tokens reassessed as attentive after retokenization.

Similarity Merging in Re-tokenization. Previous token merging methods, e.g., ToMe (Bolya et al. 2023), merge pairs of tokens with the highest similarity, selecting the fixed top-k most similar pairs per layer. However, as merging proceeds progressively, tokens with low similarity are apt to be merged, resulting in semantically essential tokens to be diluted. Tab. 6 shows a marked decline in the similarity score of ToMe from the first to the last layer during the merging procedures, indicating information loss. This loss correlates with accuracy drops of 0.47% in DeiT-S, as shown in Tab. 2, despite not discarding any tokens in ToMe.

We also found that highly attentive tokens were primarily merged, as shown in Tab. 7. For instance, in DeiT-Ti$_{r=13}$,

Model	First layer	Last layer	Gap
ToMe (Bolya et al. 2023)	0.7852	0.6763	-0.1089
ImagePiece (Ours)	0.8091	0.8096	+0.0005

Table 6: Comparison of similarity scores among merged token pairs for each sample on the ImageNet test set. We calculate the average of these scores for the 500 samples with the lowest similarity in the first and last layers.

Top-k percent (%)	30	20	15	70
DeiT-Ti$_{r=13}$	65.16	52.83	43.29	91.80
DeiT-S$_{r=13}$	60.10	48.28	39.35	88.70

Table 7: In the previous token merging method (Bolya et al. 2023), we observed the overlap between the attentive tokens and the merged tokens in the first layer.

65.16% of merged tokens are within the top-30% attentive tokens in the first layer. This suggests that important tokens, as indicated by their class attention scores, are predominantly merged, leading to potential information loss as the attentive tokens are quickly merged in the early layers. As the process progresses to later layers, the proportion of inattentive tokens increases, which may result in lower similarity during merging.

On the other hand, ImagePiece retokenizes inattentive tokens into semantically meaningful units, as shown in Tab. 6, where the similarity scores of ImagePiece remain consistent or improve by 0.0005 from the first to the last layer. Also, the similarity score of ImagePiece in the first layer is higher than that of ToMe, enabling a more effective merging process. Unlike previous token merging methods, ImagePiece discards final non-semantic tokens based on the re-evaluation after retokenization, which benefits subsequent retokenizations by eliminating truly non-semantic tokens.

During the retokenization, we identified that the similarity between inattentive tokens in the early layer was not sufficiently high, complicating the decision on which token pairs to merge, so we enhance the semantics of these inattentive tokens for effective retokenization. To improve the semantics of tokens within a visual scene through local coherence, we utilized the local coherence bias module for the retokenization to aggregate locally coherent patches. This approach tokenizes the image into tokens with overlapping patches, thereby enhancing token locality. As shown in Tab 8, ours improves the locality of tokens, which, in turn, increases the average similarity score between inattentive tokens in the first layer.

Compatibility of ImagePiece. We conducted an experiment to assess the impact of integrating our method, ImagePiece, into existing token reduction methods. Specifically, we compared the performance of the existing methods with and without ImagePiece as a tokenizer.

Following Tab. 1, we apply ImagePiece to two pruning strategies: pruning by learned projection layer (Rao et al. 2021) and pruning by [CLS] attentiveness (Liang et al. 2022). These pruning-based models, such as DynamicViT (Rao et al. 2021) and EViT (Liang et al. 2022), integrate ImagePiece into the transformer blocks, excluding the specific pruning layers, to retokenize the tokens before reevaluating them for pruning. In the merging-based method ToMe (Bolya et al. 2023) from Tab. 2, ImagePiece retokenizes non-semantic tokens into meaningful units, enabling more accurate token merging and improved performance.

Image tokenization method	Similarity score
Patch tokenizer of ViT	0.5293
Local coherence bias (Ours)	0.8091

Table 8: Comparison between the similarity scores obtained using patch tokenizer of ViTs and ours in the first layer.

Model	Acc (%)	Throughput (img/s)
DynamicViT (Rao et al. 2021)	79.32	3762.0
ImagePiece + DynamicViT	**80.11** (+0.79)	3857.1
ToMe (Bolya et al. 2023)	79.36	3806.1
ImagePiece + ToMe	**80.08** (+0.72)	3876.5
EViT (Liang et al. 2022)	79.37	3787.5
ImagePiece + EViT	**80.22** (+0.85)	3891.9

Table 9: Comparison of prior token reduction methods with and without the integration of ImagePiece on ImageNet-1k. By applying ImagePiece for retokenization, the performance of prior methods improves by an average of 0.79%.

Ablation Study. To introduce local inductive bias into visual scenes, we incorporate a *local coherence bias* module before the ImagePiece pipeline. The performance of our method without the *local coherence bias* module is 79.81%, demonstrating that the full ImagePiece framework with the *local coherence bias* is effective, as our method achieves an accuracy of 80.22%.

Conclusion

In this paper, we propose ImagePiece, a novel retokenization strategy for enhancing the efficiency of Vision Transformers. ImagePiece merges non-semantic but locally coherent tokens into meaningful chunks (or discard them if they remain irrelevant to the visual scene), significantly increasing inference speed and improving ImageNet classification accuracy. Our approach not only reduces the overall token count, but also ensures that the remaining tokens contribute more meaningfully to the overall visual understanding. Extensive experiments and analysis demonstrate that our re-tokenization approach is more effective than previous patch tokenizers used in ViTs. As for future work, this advancement sets a new standard in token management, promising significant impacts on efficient ViTs across various applications.

References

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2023. Token Merging: Your ViT But Faster. In *The Eleventh International Conference on Learning Representations*.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; and Hoffman, J. 2022. Hydra attention: Efficient attention with many heads. In *European Conference on Computer Vision*, 35–49. Springer.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, M.; Lin, M.; Li, K.; Shen, Y.; Wu, Y.; Chao, F.; and Ji, R. 2023. Cf-vit: A general coarse-to-fine method for vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7042–7052.

Choromanski, K.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Mohiuddin, A.; Kaiser, L.; et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; and Shen, C. 2021. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34: 9355–9366.

Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; and Ré, C. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35: 16344–16359.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; and Guo, B. 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12124–12134.

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

Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; and Feichtenhofer, C. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6824–6835.

Fayyaz, M.; Koohpayegani, S. A.; Jafari, F. R.; Sengupta, S.; Joze, H. R. V.; Sommerlade, E.; Pirsiavash, H.; and Gall, J. 2022. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, 396–414. Springer.

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 *International Conference on Machine Learning*, 3690–3699. PMLR.

Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; and Douze, M. 2021. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12259–12269.

Kim, G.; and Cho, K. 2020. Length-adaptive transformer: Train once with length drop, use anytime with search. *arXiv preprint arXiv:2010.07003*.

Kim, S.; Shen, S.; Thorsley, D.; Gholami, A.; Kwon, W.; Hassoun, J.; and Keutzer, K. 2022. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 784–794.

Kitaev, N.; Kaiser, Ł.; and Levskaya, A. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Kong, Z.; Dong, P.; Ma, X.; Meng, X.; Sun, M.; Niu, W.; Shen, X.; Yuan, G.; Ren, B.; Qin, M.; et al. 2022. Spvit: Enabling faster vision transformers via soft token pruning. *ECCV*.

Kudo, T.; and Richardson, J. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Lassance, C.; Maachou, M.; Park, J.; and Clinchant, S. 2021. A study on token pruning for colbert. *arXiv preprint arXiv:2112.06540*.

Li, Y.; Wu, C.-Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4804–4814.

Liang, Y.; GE, C.; Tong, Z.; Song, Y.; Wang, J.; and Xie, P. 2022. EViT: Expediting Vision Transformers via Token Reorganizations. In *International Conference on Learning Representations*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Long, S.; Zhao, Z.; Pi, J.; Wang, S.; and Wang, J. 2023. Beyond Attentive Tokens: Incorporating Token Importance and Diversity for Efficient Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10334–10343.

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

Mehta, S.; and Rastegari, M. 2022. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *ICLR*.

Meng, L.; Li, H.; Chen, B.-C.; Lan, S.; Wu, Z.; Jiang, Y.-G.; and Lim, S.-N. 2022. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12309–12318.

Michel, P.; Levy, O.; and Neubig, G. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Pan, B.; Panda, R.; Jiang, Y.; Wang, Z.; Feris, R.; and Oliva, A. 2021. IA-RED2: Interpretability-Aware Redundancy Reduction for Vision Transformers. *Advances in Neural Information Processing Systems*, 34: 24898–24911.

Pan, Z.; Cai, J.; and Zhuang, B. 2022. Fast vision transformers with hilo attention. *Advances in Neural Information Processing Systems*, 35: 14541–14554.

Pan, Z.; Zhuang, B.; He, H.; Liu, J.; and Cai, J. 2022. Less is more: Pay less attention in vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2035–2043.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C.-J. 2021. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Ryoo, M.; Piergiovanni, A.; Arnab, A.; Dehghani, M.; and Angelova, A. 2021. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in Neural Information Processing Systems*, 34: 12786–12797.

Shen, Z.; Zhang, M.; Zhao, H.; Yi, S.; and Li, H. 2021. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3531–3539.

Song, Z.; Xu, Y.; He, Z.; Jiang, L.; Jing, N.; and Liang, X. 2022. Cp-vit: Cascade vision transformer pruning via progressive sparsity prediction. *arXiv preprint arXiv:2203.04570*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021a. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021b. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 32–42.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; and Titov, I. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *ACL*.

Wang, S.; Li, B. Z.; Khabsa, M.; Fang, H.; and Ma, H. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Xiao, T.; Singh, M.; Mintun, E.; Darrell, T.; Dollár, P.; and Girshick, R. 2021. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34: 30392–30400.

Xu, Y.; Zhang, Z.; Zhang, M.; Sheng, K.; Li, K.; Dong, W.; Zhang, L.; Xu, C.; and Sun, X. 2022. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2964–2972.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yin, H.; Vahdat, A.; Alvarez, J. M.; Mallya, A.; Kautz, J.; and Molchanov, P. 2022. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10809–10818.

Yu, H.; and Wu, J. 2023. A unified pruning framework for vision transformers. *Science China Information Sciences*, 66(7): 1–2.

Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.-H.; Tay, F. E.; Feng, J.; and Yan, S. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, 558–567.

Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Jiang, Z.; Hou, Q.; and Feng, J. 2021. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*.