SMMix: Self-Motivated Image Mixing for Vision Transformers

Mengzhao Chen[1], Mingbao Lin[2], Zhihang Lin[1], Yuxin Zhang[1], Fei Chao[1], Rongrong Ji[1*]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, School of Informatics, Xiamen University
[2]Tencent Youtu Lab

Abstract

CutMix is a vital augmentation strategy that determines the performance and generalization ability of vision transformers (ViTs). However, the inconsistency between the mixed images and the corresponding labels harms its efficacy. Existing CutMix variants tackle this problem by generating more consistent mixed images or more precise mixed labels, but inevitably introduce heavy training overhead or require extra information, undermining ease of use. To this end, we propose an novel and effective Self-Motivated image Mixing method (SMMix), which motivates both image and label enhancement by the model under training itself. Specifically, we propose a max-min attention region mixing approach that enriches the attention-focused objects in the mixed images. Then, we introduce a fine-grained label assignment technique that co-trains the output tokens of mixed images with fine-grained supervision. Moreover, we devise a novel feature consistency constraint to align features from mixed and unmixed images. Due to the subtle designs of the self-motivated paradigm, our SMMix is significant in its smaller training overhead and better performance than other CutMix variants. In particular, SMMix improves the accuracy of DeiT-T/S/B, CaiT-XXS-24/36, and PVT-T/S/M/L by more than +1% on ImageNet-1k. The generalization capability of our method is also demonstrated on downstream tasks and out-of-distribution datasets. Our project is available at https://github.com/ChenMnZ/SMMix.

1. Introduction

Vision transformers (ViTs) [12] have made substantial breakthroughs across various vision tasks, such as classification [12, 44, 51, 26, 42, 3], detection [2, 60, 30, 13], and segmentation [57, 52, 22, 23]. However, the data-hungry problem [12, 44] of ViT causes a serious overfitting problem when the data is insufficient. In order to improve the gen-
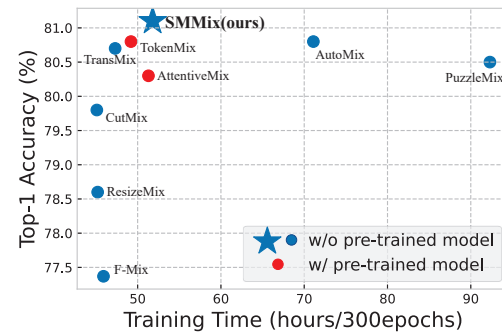

Figure 1: Training time *vs.* accuracy with DeiT-S on ImageNet-1k. SMMix outperforms existing methods with light overhead.

eralization of ViTs, data mixing augmentation techniques such as Mixup [56] and CutMix [54], are used in the ViTs training recipe. In particular, CutMix randomly crops a patch from the source image, pastes it into the target image, and forms a ground-truth label by mixing the labels of the source and target images in proportion to the area ratio of the mixed image. CutMix [54] has been demonstrated to greatly enhance the generalization of ViTs. For example, CutMix increases the top-1 accuracy of ViT-Small [12] by 4.1% [31] on ImageNet-1k [9] validation set.

Despite the progress, the image-label inconsistency issue also stems from the random patch selection and linear label combination. Figure 2 illustrates a typical example, in which the mixed image of CutMix does not contain any hints of ladybirds. However, ladybird still appears on the generated mixed label. Such an image-label inconsistency issue prevents ViTs from further improving performance. Two mainstream methods: 1) image-driven [46, 28, 48, 36] and 2) label-driven [4, 34, 26, 55, 35], have recently been considered to overcome the drawbacks of CutMix. The former method is dedicated to enhancing the saliency of mixed images, while the latter method aims to enhance the precision of mixed labels. Nevertheless, these methods usually come with heavy training overhead, such as requiring

---

*Corresponding authors: Rongrong Ji (rrji@xmu.edu.cn)

Target Image	Source Image	CutMix
ladybird: 1.0	cock: 1.0	ladybird:0.6 cock:0.4

SMMix

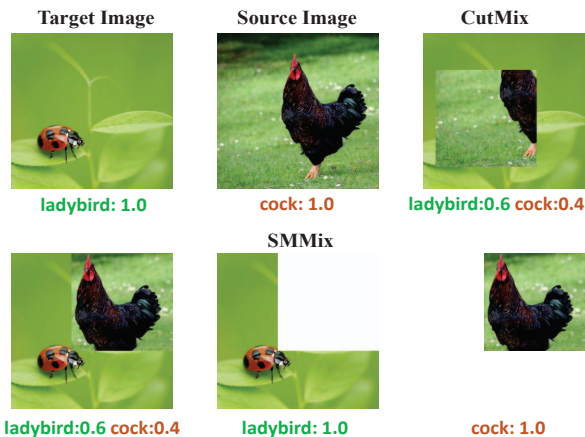| ladybird:0.6 cock:0.4 | ladybird: 1.0 | cock: 1.0 |

Figure 2: CutMix *vs.* SMMix. CutMix pastes a randomly-cropped patch in the source image to the target image, while SMMix pastes the attentive region in the source image to the unattentive region in the target image. SMMix is co-trained by three fine-grained labels instead of a simple mixed label as CutMix.

pre-trained models [48, 34, 26, 55], double forward and backward propagations [27, 28, 36], or additional generators [36], which may undermine the ease of the use of image mixing technique. Moreover, these methods only consider the image and label enhancement in isolation, resulting in limited efficiency.

To address the aforementioned challenges, we propose a novel method, **S**elf-**M**otivated image **Mix**ing (SMMix), to enhance image mixing with ViTs. By leveraging the bootstrapping capabilities of the model under training itself, SMMix simultaneously motivates image and label enhancement with light training overhead. Specially, we first use the image attention score in Eq. (6) that accumulates attention score across all the image tokens. The motivations are from a widely-accepted actuality in existing works [32, 53, 5], in which the class attention score from the self-attention operation can locate semantic objects. Therefore, we opt to use the image attention score to extend the general applicability of SMMix, since class attention is often unavailable for ViT models without a class token, while the image attention score can be easily obtained by feeding original images to a ViT model. With the guidance of the image attention score, we select the maximum-scored (most attentive) region from a source image and paste it to the region with a minimum attention score in a target image. We term this process as *max-min attention region mixing*, which alleviates the image-label inconsistency issue by enriching the attention-focused objects in mixed images.

Distinctive from the prerequisite to tuning mixed labels [4, 34], capturing attentive objects in mixed images allows for a *fine-grained label assignment*. We supervise dif-

ferent regions in a mixed image with different labels. Concretely, the output tokens of a mixed image are assigned three types of labels to accomplish the label enhancement, as illustrated in Figure 2, including mixed image label, target image label, and source image label. We aggregate all output tokens, the result is then supervised by the mixed image label. We also use region-specific supervision, *i.e.*, target image labels and source image labels, to supervise the aggregated results of tokens from the target regions and source regions, respectively.

With label-consistent mixed images, we can extract mixed image features from ViTs. To correctly recognize the mixed images, we expect the features of mixed images to fall into a consistent space with those of original unmixed images. We realize this function by creating a *feature consistency constraint*, which aligns the feature distributions between mixed images and the linear combination of unmixed images. Specially, SMMix can obtain the feature distributions and the image attention score of unmixed images from the same forward propagation of the model under training, resulting in light overhead.

Based on the above considerations, three key components are proposed in this paper, including 1) max-min attention region mixing (Sec. 4.1), 2) fine-grained label assignment (Sec. 4.2), and 3) feature consistency constraint (Sec. 4.3). We term our method *self-motivated image mixing*, since these components eliminate the dependency on pre-trained models and simply depend on the model under training itself. We have performed extensive experiments, which demonstrate the powerful ability of our SMMix to boost the performance of various ViT-based models, including DeiT [44] with a plain architecture, PVT [50] with a hierarchical architecture, CaiT [45] with deeper depth, and Swin [37] with local self-attention. Moreover, our SMMix achieves better training overhead and performance trade-off because of the self-motivated paradigm. As shown in Figure 1, our SMMix can achieve state-of-the-art top-1 accuracy with light training overhead and does not require pre-trained models.

2. Related Work

2.1. Vision Transformers

Vision Transformer (ViT) [12] shows the visual recognition ability of an original transformer [47]. However, ViT is easier to overfit on small datasets due to the lack of inductive bias. To handle this problem, DeiT [44] introduces a powerful training recipe with various data augmentations [54, 56, 8] and regularization techniques [24, 21, 58]. Based on the DeiT [44] training recipe, many ViT-based architectures [37, 42, 7, 11, 40, 20, 6, 3, 45, 50] are proposed to improve performance on various vision tasks. In this work, we focus on improving CutMix [54] augmentation,

one of the data augmentation methods in DeiT [44] training recipe.

2.2. Variants of CutMix

CutMix [54] randomly crops a patch from the source image and pastes it to the same location in the target image. Ground-truth labels of mixed images are generated by linearly combining the labels of the source and target images in proportion to the area ratio of the mixed images. However, such random crop-and-paste may cause image-label inconsistency. Existing works try to alleviate such a problem from two perspectives as follows.

Image-driven Reconstruction. Image-driven reconstruction methods aim to maximize the saliency of mixed images. SaliencyMix [46] and AttetiveMix [48] select the cropped patches based on the saliency maps, which are obtained by statistical saliency model or pre-trained model. Furthermore, based on double forward and backward propagations, PuzzleMix [28] and Co-Mixup [27] find the optimal mixed mask by solving the combinatorial optimization problem. Recently, instead of manually designing the mixing policies, AutoMix [36] trains an additional mixup generator to generate mixed samples. As can be seen, the strategies for maximizing the saliency of mixed images are becoming increasingly sophisticated. To address such problems, our SMMix proposes a simple yet effective max-min attention region mixing to enhance the mixed images.

Label-driven Reconstruction. Label-driven reconstruction methods dedicate to generating more precise labels. TransMix [4] mixes labels based on the class attention score. Other works [34, 26, 55] rely on a big-scale model pre-trained on JFT-300M [43]. Based on the activation map of pre-trained model, TokenMix [34] assigns content-based mixes labels to mixed images, TokenLabel [26] generates token-level supervision, and ReLabel [55] reorganizes the ImageNet-1k training set into a multi-label framework. Instead of depending on pre-trained models and adjusting the mixed labels, our SMMix proposes fine-grained label assignment, which provides fine-grained supervision to the output tokens by ground-truth labels.

3. Preliminary

3.1. CutMix Augmentation

CutMix [54] enhances data diversity by mixing images. Let $\mathbf{x}$ and $\mathbf{y}$ denote a training image and its label, where $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$. Given a source image-label training pair $(\mathbf{x}_A, \mathbf{y}_A)$ and a target one $(\mathbf{x}_B, \mathbf{y}_B)$, CutMix generates a new training sample $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ as follows:

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{M} \odot \mathbf{x}_A + (\mathbf{1} - \mathbf{M}) \odot \mathbf{x}_B, \\ \tilde{\mathbf{y}} &= \lambda \mathbf{y}_A + (1 - \lambda)\mathbf{y}_B, \end{aligned} \quad (1)$$

where $\mathbf{M} \in \{0, 1\}^{H \times W}$ denotes a rectangular binary mask that indicates where to drop or keep in the two images, $\odot$ is element-wise multiplication, and λ is the combination ratio sampled from a beta distribution. Note that λ indicates the area ratio of $\mathbf{x}_A$ in mixed image $\tilde{\mathbf{x}}$, *i.e.*, $\lambda = \frac{\sum \mathbf{M}}{HW}$.

3.2. Vision Transformer

Loss Computing. Given a ViT-based model $\mathcal{V}$, the output token sequence of an input image $\mathbf{x}$ is:

$$\mathcal{V}(\mathbf{x}) = [(\mathbf{X}^{cls}); \mathbf{X}^1; ...; \mathbf{X}^N], \quad (2)$$

where N is the total number of image tokens, X^i is the i-th image token, and X^{cls} corresponds to the class token, which exists only in some of the ViT-based architectures [44, 50, 45]. The final prediction distribution $\mathbf{Y}$ is obtained with a classifier $\mathcal{F}$:

$$\mathbf{Y} = \begin{cases} \mathcal{F}(\mathbf{X}^{cls}) & \text{, w/ class token,} \\ \mathcal{F}(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}^i) & \text{, w/o class token.} \end{cases} \quad (3)$$

The classification loss for the image $\mathbf{x}$ is:

$$L = CE(\mathbf{Y}, \mathbf{y}), \quad (4)$$

where $CE(\cdot, \cdot)$ represents the cross entropy function.

Self-Attention Operation. Self-attention operation is the key component of ViT. Given an image token sequence* $\mathbf{T} \in \mathbb{R}^{N \times d}$. It is firstly linearly mapped into three matrices, namely $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$. Then, the self-attention operation is computed as:

$$\begin{aligned} \mathcal{A}(\mathbf{Q}, \mathbf{K}) &= \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}) = [\mathbf{A}^1; \mathbf{A}^2; ...; \mathbf{A}^N], \\ \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \mathcal{A}(\mathbf{Q}, \mathbf{K})\mathbf{V}. \end{aligned} \quad (5)$$

The image attention score $\boldsymbol{\alpha} \in \mathbb{R}^N$ is derived as:

$$\boldsymbol{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{A}^i = [\alpha^1, \alpha^2, ..., \alpha^N]. \quad (6)$$

The image attention score above is the result of single-head self-attention. For multi-head self-attention, we simply average across all attention heads to get the final image attention score.

4. Self-Motivated Image Mixing

This section formally introduces our SMMix, a novel image mixing method that maximizes the information of the mixed image and provides more fine-grained labels. Figure 3 illustrates an overview of our proposed SMMix. Details are given below.

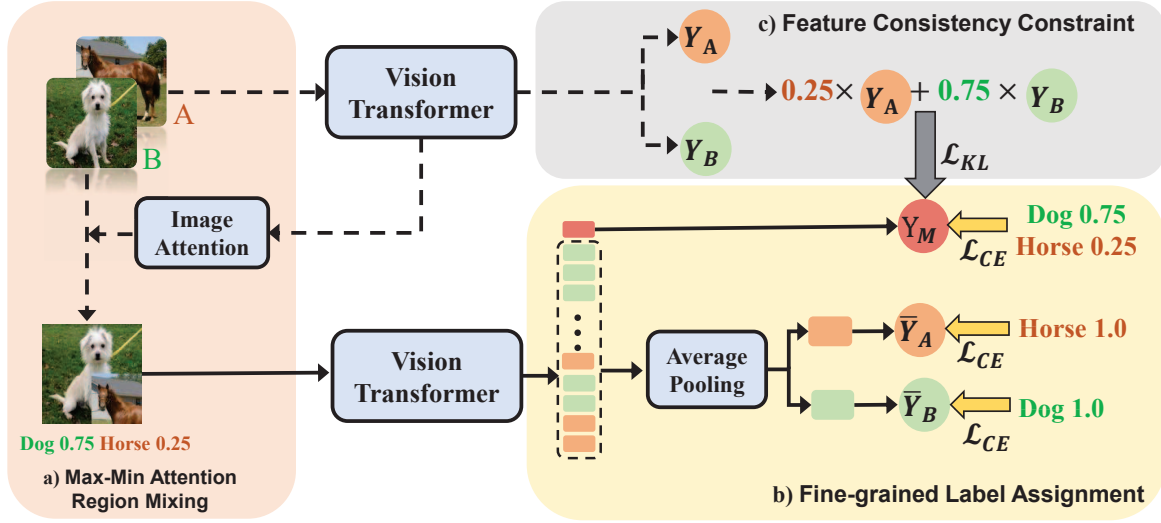*We take the case without class token as an example here.

Figure 3: Framework of SMMix that contains three components: a) max-min attention region mixing: maximizing the information of mixed images according to the image attention score. b) fine-grained label assignment: applying different supervision labels to tokens from different regions. c) feature consistency constraint: restraining the model to extract consistent features for mixed and unmixed images. The dashed arrows indicate no backward propagation. (Best viewed in colors)

4.1. Max-Min Attention Region Mixing

To maximize the information of mixed images, our max-min attention region mixing replaces a minimum-scored region of the target image with a maximum-scored region of the source image. As depicted in Figure 4, we split the source image $\mathbf{x}_A$ and the target image $\mathbf{x}_B$ into non-overlapping patches of size $P \times P$. A total of $N = \frac{H}{P} \times \frac{W}{P}$ patches are obtained for each image. Therefore, $\mathbf{x}_A$ and $\mathbf{x}_B$ are reorganized as $\mathbf{x}_A, \mathbf{x}_B \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times (P^2 C)}$, row of which corresponds to a token. Then, we feed them into a ViT model to get the corresponding image attention scores $\boldsymbol{\alpha}_A \in \mathbb{R}^N$ and $\boldsymbol{\alpha}_B \in \mathbb{R}^N$. Similarly, we rearrange the shape of their image attention score vectors, $\boldsymbol{\alpha}_A$ and $\boldsymbol{\alpha}_B$, to matrices of $\frac{H}{P} \times \frac{W}{P}$.

Similarly to CutMix, we intend to crop a region from the source image and paste the region into the target image to form a mixed image. To this effect, we introduce a side ratio δ, sampled from a uniform distribution (0.25, 0.75), to determine the total $\lfloor \delta \frac{H}{P} \rfloor \times \lfloor \delta \frac{W}{P} \rfloor$ patches within the cropped region. Our core difference in this paper is to locate the most informative region in the source image, and the least informative region in the target image. Concretely, the center indices of these two regions are defined as:

$$
i_s, j_s = \arg\max_{i,j} \sum_{p,q} \boldsymbol{\alpha}_A^{i+p-\lfloor \frac{h}{2} \rfloor, j+q-\lfloor \frac{w}{2} \rfloor},
$$

$$
i_t, j_t = \arg\min_{i,j} \sum_{p,q} \boldsymbol{\alpha}_B^{i+p-\lfloor \frac{h}{2} \rfloor, j+q-\lfloor \frac{w}{2} \rfloor}, \quad (7)
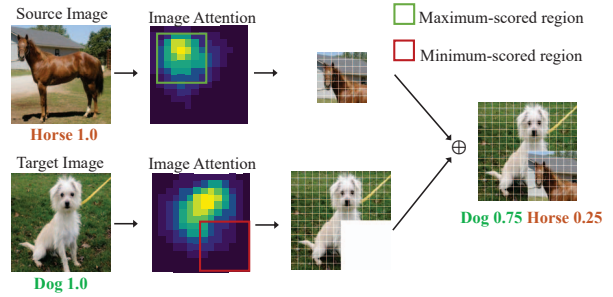$$

Figure 4: The pipeline of max-min attention region mixing.

where $h = \lfloor \delta \frac{H}{P} \rfloor$, $w = \lfloor \delta \frac{W}{P} \rfloor$, $p \in \{0, 1, ..., h-1\}$, and $q \in \{0, 1, ..., w-1\}$. It is intuitive that the selected region contains patches with the maximum attention score of the source image and the minimum attention score of the target image.

Then, in contrast to CutMix of Eq. (1), we obtain the new mixed training sample $(\mathbf{x}_M, \mathbf{y}_M)$ as follows:

$$
\mathbf{x}_M = \mathbf{x}_B,
$$

$$
\mathbf{x}_M^{i_t+p-\lfloor \frac{h}{2} \rfloor, j_t+q-\lfloor \frac{w}{2} \rfloor} = \mathbf{x}_A^{i_s+p-\lfloor \frac{h}{2} \rfloor, j_s+q-\lfloor \frac{w}{2} \rfloor},
$$

$$
\mathbf{y}_M = \lambda_M \mathbf{y}_A + (1 - \lambda_M)\mathbf{y}_B, \quad (8)
$$

where $\lambda_M = \frac{hwP^2}{HW}$.

4.2. Fine-grained Label Assignment

We feed the mixed image $\mathbf{x}_M$ to the ViT model to obtain the final output image token sequence $\mathbf{X}_M =$

$[\mathbf{X}^1; \mathbf{X}^2; ...; \mathbf{X}^N]$ and prediction distribution $\mathbf{Y}_M$. Then, the traditional classification loss is:

$$L_{cls} = CE(\mathbf{Y}_M, \mathbf{y}_M). \qquad (9)$$

Such a loss only considers the overall mixed image information. However, the introduced image mixing method (Sec. 4.1) endows objects of $\mathbf{y_A}$ and $\mathbf{y_B}$ within the content of the mixed image $\mathbf{x}_M$. Therefore, it is plausible to supervise different regions in mixed images with different labels. To achieve this purpose, we reshape the final output image token sequence $\mathbf{X}_M$ into image shape of $\mathbf{X}_M \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times d}$ where d is the final token embedding size. Accordingly, we aggregate the tokens from the source image as:

$$\bar{\mathbf{X}}_A = \frac{1}{hw} \sum_{p,q} \mathbf{X}_M^{i_t+p-\lfloor \frac{h}{2} \rfloor, j_t+q-\lfloor \frac{w}{2} \rfloor}, \qquad (10)$$

and aggregate the tokens from the target image as:

$$\bar{\mathbf{X}}_B = \frac{1}{HW - hw} \sum_{i',j'} \mathbf{X}_M^{i',j'}, \qquad (11)$$

where $(i',j') \in \left\{ (i',j') | 1 \leq i' \leq \frac{H}{P}, 1 \leq j' \leq \frac{W}{P}, (i',j') \notin \{(i_t + p - \lfloor \frac{h}{2} \rfloor, j_t + q - \lfloor \frac{w}{2} \rfloor)\} \right\}$. Then, their prediction distributions are derived from the classifier $\mathcal{F}$:

$$\begin{aligned} \bar{\mathbf{Y}}_A &= \mathcal{F}(\bar{\mathbf{X}}_A), \\ \bar{\mathbf{Y}}_B &= \mathcal{F}(\bar{\mathbf{X}}_B). \end{aligned} \qquad (12)$$

Then, SMMix supervises the fine-grained prediction distributions with fine-grained labels, $\mathbf{y}_A$ and $\mathbf{y}_B$, as:

$$L_{fine} = \frac{1}{2} \big(CE(\bar{\mathbf{Y}}_A, \mathbf{y}_A) + CE(\bar{\mathbf{Y}}_B, \mathbf{y}_B) \big). \qquad (13)$$

Such fine-grained supervision can help ViTs locate target objects and improve their recognition ability. Besides, the additional computational costs are negligible, relying only upon the existing outputs and labels.

4.3. Feature Consistency Constraint

The semantic content of the mixed image, $\mathbf{x}_M$, is equivalent to the mixing of the semantic content of the unmixed images, $\mathbf{x}_A$ and $\mathbf{x}_B$. However, the semantic content of the mixed image is more complex, increasing the difficulty of extraction of features. To help features of the mixed images fall into a consistent space with those of the original unmixed images, similar to label combination, we linearly combine the prediction distributions $\mathbf{Y}_A$ and $\mathbf{Y}_B$ of unmixed images $\mathbf{x}_A$ and $\mathbf{x}_B$, and supervise $\mathbf{Y}_M$ with the combined prediction distribution. Then, we have:

$$L_{con} = KL\big(\mathbf{Y}_M, \lambda_M \mathbf{Y}_A + (1 - \lambda_M)\mathbf{Y}_B \big), \qquad (14)$$

where $KL(\cdot, \cdot)$ represents the Kullback-Leibler divergence. Note that, the prediction distributions, $\mathbf{Y}_A$ and $\mathbf{Y}_B$ in Eq. (14), and image attention score, $\boldsymbol{\alpha}_A$ and $\boldsymbol{\alpha}_B$ in Eq. (7), of unmixed images are obtained in the same forward propagation during training. Therefore, SMMix does not rely on pre-trained models and requires only one additional forward propagation in the training process.

4.4. Training Objective

Above all, in addition to the common classification loss of Eq. (9), we also require fine-grained label assignment and feature consistency constraint losses, respective in compliance with Eq. (13) and Eq. (14). Consequently, the overall training loss of our SMMix is then written as follows:

$$L_{total} = L_{cls} + L_{fine} + L_{con}. \qquad (15)$$

5. Experiments

We evaluate SMMix in four aspects: 1) Sec. 5.1, evaluating image classification task on various ViT-based architectures, 2) Sec. 5.2, transferring pre-trained models to downstream semantic segmentation and object detection tasks, 3) Sec. 5.3, transferring pre-trained models to out-of-distribution datasets, 4) Sec. 5.4, exploring the quality of mixed images. Note that in the tables, our SMMix is highlighted in gray, and **bold** denotes the best results.

5.1. ImageNet Classification

Settings. We evaluate the ability of our SMMix to improve classification performance on ImageNet-1k dataset [9], which is a 1,000-class dataset, consisting of 1.28M training images and 50k validation images. Experiments are conducted on several recent ViT-based architectures, including DeiT [44], PVT [50], CaiT [45], and Swin [37]. All models are trained on the training set, and we report the top-1 accuracy on the validation set. For a fair comparison, we follow the implementations of the official papers. We train all models for 300 epochs. Both RandAugment [8] and Mixup [56] are used as default. We simply replace the original CutMix [54] with the proposed SMMix, and switch SMMix and Mixup with a probability of 0.5. The image attention scores $\boldsymbol{\alpha}$ in Eq. (6) are obtained from the last transformer block by feeding the unmixed images into the model under training.

Results. We first compare SMMix with recent ViT-special CutMix variants, including TransMix [4] and TokenMix [34]. As shown in Table 1, SMMix consistently surpasses TransMix ($+0.1\% \sim +1.0\%$) and TokenMix ($+0.2\% \sim +0.9\%$) in various ViT-based architectures. In particular, SMMix can boost the top-1 accuracy by more than $+1\%$ in DeiT-T/S/B [44], CaiT-XXS-24/36 [45], and PVT-T/S/M/L [50] compared with the CutMix [54] baseline. Recent TokenMix [34] also achieves 82.9% Top-1 accuracy

Model	FLOPs(G)	Top-1 Acc.(%)			
		CutMix	TransMix	TokenMix	SMMix
DeiT-T [44]	1.3	72.2	72.6	73.2	**73.6**(+1.4)
DeiT-S [44]	4.7	79.8	80.7	80.8	**81.1**(+1.3)
DeiT-B [44]	17.6	81.8	82.4	**82.9**	82.9(+1.1)
CaiT-XXS-24 [45]	2.5	77.6	-	78.0	**78.9**(+1.3)
CaiT-XXS-36 [45]	3.8	79.1	79.8	-	**80.2**(+1.1)
PVT-T [50]	1.9	75.1	75.5	75.6	**76.4**(+1.3)
PVT-S [50]	3.8	79.8	80.5	-	**81.0**(+1.2)
PVT-M [50]	6.7	81.2	82.1	-	**82.2**(+1.0)
PVT-L [50]	9.8	81.7	82.4	-	**82.7**(+1.0)
Swin-T [37]	4.5	81.2	-	81.6	**81.8**(+0.6)

Table 1: Comparison of SMMix with CutMix variants designed for ViT on ImageNet-1k. "-" indicates that the corresponding results do not report in the original paper. Blue indicates the performance improvement compared with CutMix.

Methods	DeiT-S [44]	Swin-T [37]
Vanilla [31]	75.7	80.2
CutMix [54]	79.8	81.2
AttentiveMix [48]	80.3	81.3
SaliencyMix [46]	79.9	81.4
PuzzleMix [28]	80.5	81.5
F-Mix [15]	77.4	79.6
ResizeMix [39]	78.6	81.4
AutoMix [36]	80.8	**81.8**
SMMix (Ours)	**81.1**	**81.8**

Table 2: Comparison of SMMix with other CutMix variants on ImageNet-1k. We get the performance of previous CutMix variants on DeiT-S and Swin-T from the Open-Mixup [31] benchmark.

Backbone	mIoU(%)	mAcc(%)
PVT-T [50]	36.6	46.7
SMMix-PVT-T	**37.3**(+0.7)	**48.1**(+1.4)
PVT-S [50]	41.9	53.0
SMMix-PVT-S	**43.0**(+1.1)	**54.1**(+1.1)

Table 3: Transferring the pre-trained models to downstream semantic segmentation task using Semantic FPN with PVT backbone on ADE20K dataset.

Backbone	AP^b	AP^b_{50}	AP^b_{75}
PVT-T [50]	36.7	59.2	39.3
SMMix-PVT-T	**37.1**(+0.4)	**59.8**(+0.6)	**39.6**(+0.3)
PVT-S [50]	40.4	62.9	43.8
SMMix-PVT-S	**41.0**(+0.6)	**63.9**(+1.0)	**44.4**(+0.6)

Table 4: Transferring the pre-trained models to downstream object detection task using Mask R-CNN with PVT backbone on COCO val2017 dataset.

with DeiT-B, but it re-quires a pre-trained NFNet-F6 model with 438M parameters. For models with stronger inductive bias, such as Swin-T, SMMix also provides +0.6% performance improvement.

In Table 2, we further compare SMMix with other CutMix variants, including AttentiveMix [48], SaliencyMix [46], PuzzleMix [28], F-Mix [15], ResizeMix [39], and AutoMix [36]. Observably, SMMix has a performance advantage over other methods. Note that SMMix is also less overhead than previous methods that require pre-trained models [48, 26], double forward and backward propagations [28], or additional generators [36]. Specially, AutoMix [36] has the same performance as our SMMix in Swin-T. However, AutoMix requires more training time (See Figure 1 for detail) since AutoMix requires training an additional generator.

5.2. Downstream Tasks

To verify the generalization of our method, we also evaluate our SMMix pre-trained models on downstream tasks, including semantic segmentation and object detection. PVT [50] is selected as the backbone, and we follow all training settings on PVT [50] for fair comparisons.

Semantic segmentation. We use ADE20K [59] to evaluate the performance of semantic segmentation task. ADE20k is a challenging scene parsing dataset covering 150 semantic categories, with 20k, 2k, and 3k images for training, validation, and testing. We evaluate PVT backbones with Semantic FPN [29]. As shown in Table 3, SMMix improves PVT-T for +0.7% mIoU and PVT-S for +1.1% mIoU.

Object detection. We choose the challenging COCO benchmark [33] for the object detection task. All models are trained on COCO train2017 (118k images) and evaluated on val2017 (5k images). We evaluate PVT backbones with Mask R-CNN [16]. Table 4 shows that SMMix improves PVT-T for +0.4% box AP, and PVT-S for +0.6% box AP.

These results demonstrate that the models pre-trained with the proposed SMMix consistently improve the performance on downstream tasks. Therefore, SMMix can be widely used for model training because of its excellent generalization. Note that not all augmentation-based pre-training methods bring benefits to downstream tasks. For example, CutMix [7] has observed that pre-training with Mixup [56] and CutOut [10] failed to improve the object detection performance over the vanilla pre-trained models.
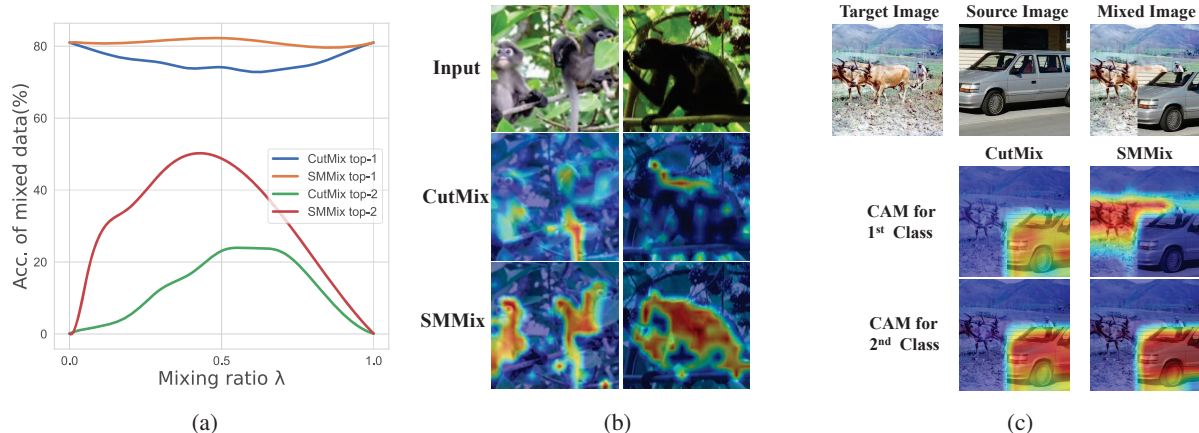
Figure 5: (a) Top-1/2 accuracy of mixed images on ImageNet-1k. Top-1 accuracy is calculated by counting the top-1 prediction belongs to $\{y_A, y_B\}$. Top-2 accuracy is calculated by counting the top-2 prediction equal to $\{y_A, y_B\}$. (b) and (c) show the class activation map [41] of the models trained with CutMix and SMMix by testing on unmixed and mixed images, respectively.

| Model | CutMix/SMMix Top-1 Acc.(%) | | | |
	ImageNet-A	Rendition	Sketch	Stylized
DeiT-T	7.1/**8.5**	33.2/**34.9**	20.3/**22.2**	10.7/**11.2**
DeiT-S	18.7/**22.0**	42.5/**43.9**	29.5/**31.2**	15.2/**16.6**
DeiT-B	25.2/**28.1**	50.2/**51.7**	36.3/**38.1**	21.5/**22.3**
PVT-T	7.7/**9.4**	34.1/**35.2**	21.3/**22.2**	11.7/**12.5**
PVT-S	17.7/**20.4**	40.5/**41.8**	27.1/**29.2**	13.8/**15.4**
PVT-M	24.8/**28.3**	42.1/**44.4**	30.1/**31.4**	13.3/**15.6**
PVT-L	26.3/**30.0**	44.1/**44.9**	29.9/**31.5**	14.0/**16.3**
Swin-T	20.7/**22.3**	41.8/**43.1**	29.2/**29.5**	13.5/**13.8**

Table 5: Performance of various ViT architectures trained with CutMix/SMMix on ImageNet-1k and evaluated on four out-of-distribution datasets. Acc1/**Acc2** refers to the top-1 accuracy of the models trained with CutMix and SM-Mix, respectively.

Model	Label Reconstruction	Top-1 Acc.(%)
	TransMix [4]	81.1
DeiT-S [44]	TokenMix [34]	**81.2**
	w/o (ours)	81.1

Table 6: The performance of DeiT-S when introducing label reconstruction methods into SMMix.

5.3. Robustness

To verify whether SMMix can improve the robustness of ViT-based models, we also evaluate our SMMix on four out-of-distribution datasets: (i) ImageNet-A [19] contains 7,500 adversarial examples for 200 ImageNet classes, which would yield low-confidence predictions with ResNet-50 [17]. (ii) ImageNet-Rendition [18] contains 30,000 im-

age renditions (*e.g.* paintings, sculpture) for 200 ImageNet classes. (iii) ImageNet-Sketch [49] consists of sketch-like images that match the ImageNet-1k validation set in terms of category and scale. (iv) ImageNet-Stylized [14] is created by applying AdaIN [25] style transfer to ImageNet images. We train all models on ImageNet-1k [9] training set and test them on the above out-of-distribution datasets. Table 5 shows that the proposed SMMix can have consistent performance gains over CutMix on the out-of-distribution data. Such results demonstrate that SMMix can enhance the robustness of the ViT-based models.

5.4. Performance Analysis

Premium Mixed Images. The image-label inconsistency issue hinders further performance improvement of CutMix. To solve this problem, our SMMix proposes max-min attention region mixing technique, which maximizes the attentive objects in mixed images. Following AutoMix [36], we statistic the top-1/2 accuracy to verify the quality of mixed images. As shown in Figure 5a, our SM-Mix significantly improves the top-1/2 accuracy of mixed images compared with CutMix. Especially for the top-2 accuracy, our SMMix achieves 48.3% while CutMix only reaches 23.8%. Such a substantial performance improvement demonstrates that SMMix can enrich discriminative features in mixed images. To further verify the quality of mixed images generated by SMMix, we also introduce the recent label-driven reconstruction techniques [4, 34] into SMMix. Table 6 shows that the label reconstruction methods bring negligible performance improvement, +0% for TransMix [4] and +0.1% for TokenMix [34]. These results demonstrate that the max-min attention region mixing technique successfully alleviates the image-label inconsistency

Max-Min Attention Image Mixing	✗	✓	✗	✓	✓	✓
Fine-grained Label Assignment	✗	✗	✗	✓	✗	✓
Feature Consistency Constraint	✗	✗	✓	✗	✓	✓
Top-1 Acc.(%)	79.8	80.4	80.3	80.9	80.8	**81.1**

Table 7: Ablation of each component of SMMix on ImageNet-1k with DeiT-S.

problem by maximizing the information of mixed images. In general, SMMix generates better-quality training samples to help further improve performance.

Visualization. In Figure 5b and Figure 5c, we visualize the class activation map [41] of the models trained with Cut-Mix and SMMix. Note that we choose images that can be correctly classified by both the CutMix and SMMix models. It can be seen in Figure 5b that the SMMix model can locate objects more accurately than the CutMix model in the unmixed images. Furthermore, Figure 5c shows that for the mixed images, the SMMix model can accurately locate objects from two different images. On the contrary, the Cut-Mix model focuses only on the cropped regions. The misplacement of CutMix is due to the fact that the cropped regions with sharp rectangle boundaries enhance first/second-order feature statistics, resulting in self-attention operation generating basic attention scores for cropped regions regardless of content [4]. However, our SMMix provides fine-grained supervision for tokens from different regions, which can help the model locate the correct region. In the supplementary material, we also present statistical results of image attention scores that quantitatively demonstrate the phenomena observed by visualization.

5.5. Ablation studies

In this section, we conduct various ablation studies to better understand SMMix. We use DeiT as the backbone, with the same training settings as described in Sec. 5.1 unless otherwise specified.

Necessity of each design. We first analyze the efficacy of each design in our SMMix. Note that the fine-grained label assignment must be used in conjunction with the max-min attention region mixing. In Table 7, we increasingly add each component to the vanilla DeiT-S training recipe, where ✓ and ✗ denote whether or not the corresponding component is enabled. Observably, each designed component can improve the final performance. Hence, the three designs are critical to the final performance of our SMMix.

Side ratio δ of cropped rectangle. δ determines the size of the cropped rectangle, which indicates the strength of regularization. We test three strategies: 1) fixed as 0.5, 2) sampled from $U(0.25, 0.75)$, 3) sampled from $U(0,1)$. Table 8 shows that three strategies achieve similar performance, which means that the proposed SMMix is robust to the side ratio δ. We simply sample δ from the uniform dis-

Model	δ	Top-1 Acc.(%)
	0.5	81.0
DeiT-S [44]	$U(0,1)$	81.0
	$U(0.25,0.75)$	**81.1**

Table 8: Ablation of side ratio δ. U means uniform distribution.

Model	Motivated	Pre-trained	Top-1 Acc.(%)
	DeiT-T	✓	73.5
DeiT-T [44]	DeiT-S	✓	**74.1**
	self (ours)	✗	73.6

Table 9: The performance of DeiT-T with different motivate models. The self means SMMix proposed in this paper.

tribution (0.25,0.75) by default.

Pre-trained models. An essential advantage of our work is that SMMix relies entirely on the model under training itself, *i.e.*, no extra pre-trained models are required. Specially, SMMix first forwards the unmixed images to obtain the corresponding image attention score and prediction distribution for guiding the formal training process. We consider the forwarded model before formal training as a motivated model. Therefore, whether a pre-trained motivated model can provide better guidance than the model under training remains a question. For this purpose, we train DeiT-T with three motivated models: the model under training (self), pre-trained DeiT-T, and pre-trained DeiT-S. Table 9 shows that a larger-scale pre-trained model can further improve performance, while a pre-trained model on the same scale as the model under training does not provide any benefit. Thus, it is noteworthy that we use the self-motivated paradigm without pre-train models for light training overhead. However, the proposed training technique can perform better with a larger pre-trained model, which demonstrates the potential of SMMix.

6. Conclusion

This paper proposes SMMix, a novel and effective image mixing technique. Specially, we design a self-motivated paradigm that motivates both the image and label enhancement in image mixing by the model under training itself. Thus, SMMix is more flexible and easier to use than the ex-

isting CutMix variants because it has a light training overhead and eliminates the reliance on pre-trained models. Extensive experiments verify the generalization and effectiveness of SMMix, which can significantly improve the performance of various ViT-based models. Besides, SMMix also exhibits transferability on downstream tasks and robustness to out-of-distribution datasets. Overall, we hope that the self-motivated paradigm introduced by SMMix can provide a new perspective on image mixing techniques and even on deep neural network training.

Limitation. We further discuss unexplored limitations, which will be our future focus. First, SMMix somewhat increases the training overhead compared with vanilla CutMix due to the need for extra forward propagation. Second, SMMix is based on the self-attention and patch-splitting operation of ViTs. More efforts can be made to transfer the idea of SMMix to convolutional neural networks.

Acknowledgement

References

[1] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning (ICML)*, pages 1059–1071, 2021. 12

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229, 2020. 1

[3] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *International Conference on Computer Vision (ICCV)*, pages 357–366, 2021. 1, 2

[4] Jie-Neng Chen, Shuyang Sun, Ju He, Philip HS Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12135–12144, 2022. 1, 2, 3, 5, 7, 8, 13

[5] Mengzhao Chen, Mingbao Lin, Ke Li, Yunhang Shen, Yongjian Wu, Fei Chao, and Rongrong Ji. Cf-vit: A general coarse-to-fine method for vision transformer. *arXiv preprint arXiv:2203.03821*, 2022. 2

[6] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *International Conference on Computer Vision (ICCV)*, pages 589–598, 2021. 2

[7] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9355–9366, 2021. 2, 6

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Computer Vision and Pattern Recognition (CVPR) workshops*, pages 702–703, 2020. 2, 5

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *International Conference on Computer Vision (ICCV)*, pages 248–255, 2009. 1, 5, 7, 12

[10] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 6

[11] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 12124–12134, 2022. 2

[12] 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. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2

[13] Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. Adamixer: A fast-converging query-based object detector. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5364–5373, 2022. 1

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. 7

[15] Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, Adam Prügel-Bennett, and Jonathon Hare. Fmix: Enhancing mixed sample data augmentation. *arXiv preprint arXiv:2002.12047*, 2020. 6

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 7

[18] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021. 7

[19] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In

Computer Vision and Pattern Recognition (CVPR), pages 15262–15271, 2021. 7

[20] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *International Conference on Computer Vision (ICCV)*, pages 11936–11945, 2021. 2

[21] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8129–8138, 2020. 2

[22] Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Yan Wang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. Istr: End-to-end instance segmentation with transformers. *arXiv preprint arXiv:2105.00637*, 2021. 1

[23] Jie Hu, Linyan Huang, Tianhe Ren, Shengchuan Zhang, Rongrong Ji, and Liujuan Cao. You only segment once: Towards real-time panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17819–17829, 2023. 1

[24] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision (ECCV)*, pages 646–661, 2016. 2

[25] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 7

[26] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 3, 6, 12

[27] Jang-Hyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-mixup: Saliency guided joint mixup with supermodular diversity. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3

[28] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning (ICML)*, pages 5275–5285, 2020. 1, 2, 3, 6

[29] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6399–6408, 2019. 6

[30] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13619–13627, 2022. 1

[31] Siyuan Li, Zedong Wang, Zicheng Liu, Di Wu, and Stan Z Li. Openmixup: Open mixup toolbox and benchmark for visual representation learning. *arXiv preprint arXiv:2209.04851*, 2022. 1, 6

[32] 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 reorganiza-

tions. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 6

[34] Jihao Liu, Boxiao Liu, Hang Zhou, Hongsheng Li, and Yu Liu. Tokenmix: Rethinking image mixing for data augmentation in vision transformers. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 5, 7

[35] Zicheng Liu, Siyuan Li, Ge Wang, Cheng Tan, Lirong Wu, and Stan Z Li. Decoupled mixup for data-efficient learning. *arXiv preprint arXiv:2203.10761*, 2022. 1

[36] Zicheng Liu, Siyuan Li, Di Wu, Zhiyuan Chen, Lirong Wu, Jianzhu Guo, and Stan Z Li. Automix:unveiling the power of mixup for stronger classifiers. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 6, 7

[37] 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 *International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 2, 5, 6

[38] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017. 13

[39] Jie Qin, Jiemin Fang, Qian Zhang, Wenyu Liu, Xingang Wang, and Xinggang Wang. Resizemix: Mixing data with preserved object information and true labels. *arXiv preprint arXiv:2012.11101*, 2020. 6

[40] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10853–10862, 2022. 2

[41] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 7, 8, 14

[42] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. *arXiv preprint arXiv:2205.12956*, 2022. 1, 2

[43] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *International Conference on Computer Vision (ICCV)*, pages 843–852, 2017. 3

[44] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357, 2021. 1, 2, 3, 5, 6, 7, 8, 13

[45] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *International Conference on Computer Vision (ICCV)*, pages 32–42, 2021. 2, 3, 5, 6

[46] AFM Uddin, Mst Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae, et al. Saliencymix: A saliency guided data augmentation strategy for better regularization. *arXiv preprint arXiv:2006.01791*, 2020. 1, 3, 6, 12

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 2

[48] Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 3, 6

[49] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 7

[50] 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 prediction without convolutions. In *International Conference on Computer Vision (ICCV)*, pages 568–578, 2021. 2, 3, 5, 6

[51] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[52] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12077–12090, 2021. 1

[53] 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 *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 2

[54] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019. 1, 2, 3, 5, 6

[55] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2340–2350, 2021. 1, 2, 3

[56] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1, 2, 5, 6

[57] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6881–6890, 2021. 1

[58] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 13001–13008, 2020. 2

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Computer Vision and Pattern Recognition (CVPR)*, pages 633–641, 2017. 6

[60] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021. 1