TokenMix: Rethinking Image Mixing for Data Augmentation in Vision Transformers

Jihao Liu[1,2], Boxiao Liu[3], Hang Zhou[1], Hongsheng Li[1✉], and Yu Liu[2✉]

[1] CUHK, MMLab
[2] SenseTime Research
[3] SKLP, Institute of Computing Technology, CAS

Abstract. CutMix is a popular augmentation technique commonly used for training modern convolutional and transformer vision networks. It was originally designed to encourage Convolution Neural Networks (CNNs) to focus more on an image's global context instead of local information, which greatly improves the performance of CNNs. However, we found it to have limited benefits for transformer-based architectures that naturally have a global receptive field. In this paper, we propose a novel data augmentation technique TokenMix to improve the performance of vision transformers. TokenMix mixes two images at token level via partitioning the mixing region into multiple separated parts. Besides, we show that the mixed learning target in CutMix, a linear combination of a pair of the ground truth labels, might be inaccurate and sometimes counter-intuitive. To obtain a more suitable target, we propose to assign the target score according to the content-based neural activation maps of the two images from a pre-trained teacher model, which does not need to have high performance. With plenty of experiments on various vision transformer architectures, we show that our proposed TokenMix helps vision transformers focus on the foreground area to infer the classes and enhances their robustness to occlusion, with consistent performance gains. Notably, we improve DeiT-T/S/B with +1% ImageNet top-1 accuracy. Besides, TokenMix enjoys longer training, which achieves 81.2% top-1 accuracy on ImageNet with DeiT-S trained for 400 epochs.

Keywords: Data augmentation, representation learning

1 Introduction

Deep neural networks dominate the learning of visual representations and show effectiveness on various downstream tasks, including image classification [9, 11], object detection [18], semantic segmentation [37], etc. To further improve the performance, various data augmentation strategies were introduced, including hand-crafted [34, 32] and automatically searched ones [7, 8]. Recently, data augmentation based on mixing multiple images into single ones shows impressive performances on various vision tasks. The labels of such "mixed" images are created based on their original labels. Mixup [34] for the first time attempted to generate mixed training samples via linear combinations of pairs of samples.

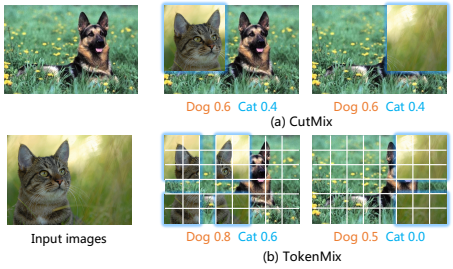


Fig. 1: **TokenMix** and CutMix. TokenMix not only mixes images at token-level to encourage better learning of long-range dependency but also generates more reasonable target scores according to content-based neural activation maps from an (even imperfectly trained) teacher network.

CutMix [32] proposed to mix pairs of samples on region level, which replaces a random local rectangular area in a source image with the contents of the corresponding area in a target image. In addition, a series of works attempted to improve CutMix with more complicated strategies on choosing rectangular sizes and locations to be used for mixing [29, 16, 15].

In general, CutMix and its variants use the region-level cut-and-paste mixing technique to enforce Convolution Neural Networks (CNNs) to pay more attention to the image's global context instead of just local information. While the CutMix augmentation can also be used for training vision transformers [11, 25], the region-level mixing strategy becomes less effective.

We revisit the design of CutMix augmentation and argue that it is a sub-optimal strategy for transformer-based architectures. On the one hand, the region-level mixing in CutMix cuts a rectangular area in a source image and mixes the contents into a target image. As CNNs are primarily designed to encode local image contents, the region-level mixing of CutMix can effectively prevent CNNs from over-focusing on local context. However, for transformer-based architectures that naturally have global receptive fields from the first layer, the region-level mixing is less beneficial. On the other hand, CutMix assigns a mixed label for the augmented image according to only the cropped area ratio between the source and target images, regardless of their cropped contents. However, the cut region and location of CutMix are randomly chosen, and the same label is assigned no matter whether the cut contents are foreground or background, which inevitably introduces label noise to the learning targets and causes unstable training (see Fig. 1(b)). There are recent works trying to mitigate this problem by attentively choosing the salient area for cutting [29, 27] or using alternate optimization to determine the cutting region [16, 15]. However, the label noise problem is still under-explored as the salient areas might not correctly correspond to the foreground regions.

In this paper, we propose TokenMix, a token-level augmentation technique that can be well applied to training various transformer-based architectures. In

contrast to previous approaches, TokenMix directly mixes two images at the token level to promote the interactions of the input tokens and generates a more reasonable target with considering the images' semantic information. First, to train transformers for better encoding long-range dependency, we directly cut at token level and allow the cut region to be separated into multiple isolated parts. As a result, the cut area can be distributed all over the image, as shown in Figure 1 (b). The token-level mixing encourages the transformer to better encode long-range dependency to correctly classify the mixed images with the augmented tokens inside. Instead of relying on alternate optimization or an extra network to determine which region to mix, all the mixing tokens in TokenMix are randomly determined as blocks, which is easier to implement with a small number of hyper-parameters.

Besides, previous methods usually assign a mixed target to the augmented image, which equals the linear combination of the ground truth labels of the source and target images. The linear combination ratio of the labels is determined as the area ratio between the cutting region of the source image and the total size of the target image. We found that such target scores can be highly inaccurate. As shown in Fig. 1 (a), the same target is assigned to both cases even the mixing area has significantly different semantic meanings.

Following the spirit of distillation, we propose to assign the target score to an augmented target image according to the content-based neural activation maps of the two mixing images. Specifically, we first obtain the neural activation maps of both the source and target images with a pre-trained neural network, which does not need to be perfectly trained. The scores of two mixing regions are calculated as the summation of the spatially normalized neural activation maps, which are combined as the final target. Our intuition is that the neural activation map of even a partially trained classification network can better localize some part of an object [36, 4] than using naive score averaging. After spatial normalization of the neural activation map, the regions with rich semantic information would be assigned high scores and low scores would be assigned for other regions, leading to more robust targets. The neural activation maps are generated offline, so the extra training overhead introduced is negligible (+0.8%). In contrast, the distillation method used in DeiT [25] relies on the online inference of a teacher network to generate target scores from augmented images, which cannot generate target scores offline and therefore nearly doubles the training time. Although ReLabel [33] and TokenLabeling [14] also explored utilizing neural activation maps to generate training supervisions, their approaches use the patch-level activations as supervisions and is prone to suffer more from inaccurate activation maps of mix-based augmentations. In contrast, our proposed method sums up the activations from the cut regions as the image-level target scores and is less likely to be affected by individual tokens' incorrect activations. Experiments on combining our token cutting strategy and ReLabel or TokenLabeling validate our scoring strategy. We show that the resulted targets of our approach are more reasonable, which improve the performance and stabilize the training of not only our proposed TokenMix and also the original CutMix. Replacing the way to generate

target scores in CutMix with our approach, we obtain a +0.7% top-1 accuracy gain on ImageNet with DeiT-S. In addition, as the generated target scores are more learning-friendly, we show that our approach enjoys longer training. Specifically, we achieve 81.2% top-1 accuracy on ImageNet with DeiT-S when training for 400 epochs.

In summary, our contributions are as follows:

- We propose TokenMix, a token-level augmentation technique that generalizes well across various transformer-based architectures.
- We propose to assign the target scores of the mixed images with content-based neural activation maps, which can benefit both TokenMix and CutMix augmentations.
- Experimental results show that TokenMix promotes transformer's capability on encoding image contents and robustness to the occlusions. We improve DeiT-S from 79.8% to 80.8% top-1 accuracy on ImageNet.

2 Related Works

Cutting-based data augmentation. The motivation behind cutting-based methods [10, 35, 23, 6] is to make a network learn informative representations from the entire image. By masking some areas from the input image, it can alleviate the issue of overfitting and improve the occlusion robustness [10]. Cutout [10] is a pioneer of this idea, and proposes to randomly select a square patch of an image and set the inputs within as some consistent. The shape and size of the masked patch are manually designed. Random-erasing [35] works in a similar way with Cutout, but introduces more randomness into the augmentation. In every iteration, the erasing operation is performed under a probability, and the size and aspect ratio is randomly selected with predefined limits. Hide-and-seek [23] differs from the previous two methods in the number of masked patches. It divides an image into grids and masked each grid randomly and independently.
Mixing-based data augmentation. Mixing-based data augmentations [34, 28, 15, 13] is another popular regularization method to help the optimization of deep neural networks. Mixup [34] proposes to mix the RGB values of two randomly selected images according to a mixing factor, which is drawn from a beta distribution. The target for the mixed image is also a linear combination of the targets of original images. Manifold Mixup [28] extends the mixed information from input images to intermediate feature maps of a network. Co-Mixup [15] and Puzzle Mix [16] consider the mixing process as an optimization problem, and propose to maximize the saliency in the mixed images. AugMix [13] generates mixed images from the original image and its transformed ones.
Joint of cutting and mixing. One issue of cutting-based augmentation is the information in the cut area is lost, so recent researches [32, 24, 22, 5] propose to combine cutting and mixing together to achieve better performance. As introduced in CutMix, a patch is replaced with that from another image instead of being deleted. Like Mixup, the target for the mixed image is computed as the proportion of the replaced area. Attentive CutMix [29] points out that the

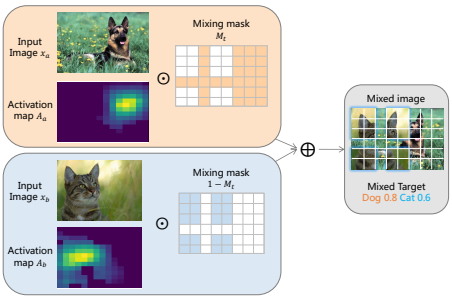


Fig. 2: **The overall pipeline of TokenMix.** TokenMix partitions the mask region into multiple separated parts. The target score of the mixed image is calculated according to the neural activation maps of the two input images.

randomly selected patches may contain only background regions, and proposes to replace attentive regions identified by a pre-trained network. RICAP [24] introduces another way of stitching four rectangle patches from different images into one new image. The target of the new image is also determined according to the area of different patches. ResizeMix [22] argues the traditional cut-and-paste operation may lead to an unreasonable target when only the background part of one image is mixed. It solves this issue by using a resize-and-paste process. In this paper, we revisit the CutMix method for vision transformer and find that CutMix under-explores the ability of vision transformer to model long-range interaction and the assigned target is non-optimal. We further introduce our TokenMix augmentation with novel ways to select mixed parts and generate learning targets.

3 Method

In this section, we first revisit the general process of CutMix [32] and show the limitations of applying CutMix to transformers. We then present our proposed TokenMix, which conducts image augmentation via mixing images at token-level and assigns target scores with neural activation maps.

3.1 Revisiting CutMix Augmentation

To enhance the localization ability of CNNs, CutMix [32] proposed to mix pairs of samples with a random rectangular binary mask. Let $x \in \mathbb{R}^{H \times W \times C}$ and y denote a training image and its label, respectively. Given a pair of training samples (x_a, y_a) and (x_b, y_b), CutMix generates a new training sample $(\tilde{x}, \tilde{y})$ as follows:

$$\begin{aligned} \tilde{x} &= M \odot x_a + (1 - M) \odot x_b, \\ \tilde{y} &= \lambda y_a + (1 - \lambda) y_b, \end{aligned} \tag{1}$$

where $M \in \{0, 1\}^{H \times W}$ denotes the rectangular mask that decides where to drop out and fill in the contents of the two images, $\odot$ denotes element-wise

multiplication, and λ is sampled from a beta distribution Beta(α, α). The binary mask M is a randomly sampled rectangle, which guarantees $\frac{\sum M}{HW} = \lambda$. Similar to Mixup [34], CutMix assigns a mixed target for the generated image as a linear combination of y_a and y_b.

We argue that the region-level mixing in CutMix might not be suitable for transformer-based architectures. As CNNs are primarily designed to encode local image contents, training with CutMix effectively prevents CNNs from over-focusing on local context. However, transformer-based architectures might be less benefited from CutMix as all of its layers have global receptive fields. In addition, the label of the mixed image is a linear combination of y_a and y_b with mixing ratio λ being estimated only according to the size of the mask, which might be inappropriate in many cases as shown in Figure 1 (b). Although there were recent methods on attempting to improve CutMix by choosing the salience regions to maximize the saliency in the mixed images [15, 16, 29, 27], the salient areas might not correctly correspond to the target class [2], and the label noise problem is still serious.

3.2 TokenMix

In this paper, we propose TokenMix to mix a pair of images to generate a mixed image and learning target. We generate the mask M at token-level to encourage better learning of long-range dependency and assign the target score of the mixed image according to the content-based neural activation maps of the two mixing images, which follows the general spirit of distillation to create more robust targets.

Figure 2 shows an overview of our proposed TokenMix. We first partition the input image x into non-overlapping patches $x^p \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times (P^2 \cdot C)}$, which are then linearly projected to visual tokens. We then generate a random mask $M_t \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P}}$ at token-level according to the mask-out ratio λ. The mixed new training sample $(\tilde{x}^p, \tilde{y})$ is created as follows:

$$\begin{aligned} \tilde{x}^p &= M_t \odot x_a^p + (1 - M_t) \odot x_b^p, \\ \tilde{y} &= \sum_{i \in \mathfrak{S}} M_{ti} \odot A_{ai} + \sum_{i \in \mathfrak{S}} (1 - M_{ti}) \odot A_{bi}, \end{aligned} \tag{2}$$

where $\mathfrak{S}$ indicates the set of all tokens, $\odot$ denotes element-wise multiplication, M_{ti} denotes the i-th token of the mask M_t, A_{ai} and A_{bi} are the i-th token of the spatially normalized neural activation maps of x_a and x_b respectively. The neural activation maps are generated with pre-trained networks' last layer before the classification head [14, 33].

Instead of masking a whole rectangular area, we partition the mask area into multiple separated parts. For each part, we randomly choose the number of masked tokens and the aspect ratio [1, 32]. We set the minimum number of tokens to 14 and log-uniformly sample the aspect ratio in the range of $[0.3, \frac{1}{0.3}]$. We repeatedly mask a part of the image until the total number of masked tokens reaches the pre-defined ratio $\lambda \frac{HW}{P^2}$. Instead of sampling λ from a beta

distribution, we set λ to 0.5 unless otherwise specified. Our intuition is that the distributed masking regions are easier to recognize compared with masking a whole rectangular area. For investigation, we also introduce a uniformly random version, where each masking part is only a single token. While totally random mixing is harmful to the performance of CNNs, we show that transformers are still benefited from the simplified version.

To solve the issue of inaccurate target scores generated by CutMix, we propose to set the target score with the content-based neural activation maps of the two mixing images, generated by a pre-trained teacher network. Our intuition is that not all regions correspond to the foreground object. Concretely, the regions with rich semantic information would have a bigger impact on the target score than other regions. Inspired by the distillation technique that sets the target score of an image by a teacher network, we extend the design to set the target score by combining a teacher network's neural activation maps of the two mixing images. As shown in Figure 2, the target scores of two mixing regions are calculated as the summation of spatially normalized neural activation maps within the mask for x_a or outside the mask for x_b. We then combine the two target scores as the final target of the mixed image.

Compared to previous arts [32, 29, 16, 15], our proposed TokenMix has two main advantages: 1) We explicitly encourage the transformer to better encode long-range dependency to correctly classify the image with the other image mixed inside. We show that our approach can lead to consistent accuracy gain when used in various vision transformers, and also enhances the occlusion robustness of the transformers. 2) The target label of the mixed image that is generated with content-based neural activation map is more robust than those of previous approaches, which takes advantage of the distillation technique. Besides, we show that our approach promotes transformers to better localize the discriminate regions, with attention weights.

4 Experiment

4.1 Datasets

We use ImageNet-1K [9] dataset to demonstrate the effectiveness of our method. The dataset contains 1.2 million images for training and 50K for validation. The top-1 accuracy is reported as the evaluation metric.

We also use ADE20K [37] to verify the transferability of our TokenMix pre-trained models. ADE20K is a widely-used semantic segmentation dataset, covering 150 semantic categories. The dataset has 25K images in total, with 20K for training, 2K for validation, and another 3K for testing.

4.2 Implementation Details

We evaluate our method on several recent vision transformer architectures, including DeiT [25], CaiT [26], PVT [30] and Swin Transformer [30]. We also test

Table 1: ImageNet classification performances based on various transformer-based architectures. TokenMix consistently improves DeiT for $\sim 1\%$ top-1 accuracy with nearly no extra training overhead.

Model	#FLOPs (G)	#Params (M)	CutMix	TokenMix
DeiT-T [25]	1.3	5.7	72.2	**73.2** (+1.0)
PVT-T [30]	1.9	13.2	75.1	**75.6** (+0.5)
CaiT-XXS-24 [26]	2.5	9.5	77.6	**78.0** (+0.4)
DeiT-S [25]	4.6	22.1	79.8	**80.8** (+1.0)
Swin-T [19]	4.5	29	81.2	**81.6** (+0.4)
DeiT-B [25]	17.6	86.6	81.8	**82.9** (+1.1)

TokenMix on ResNet [12], which is representative of convolution models, as comparison. We follow the training recipe of DeiT [25]. The batch size is set to 1024. We use AdamW [17, 20] as the optimizer and set the learning rate as 0.001 with 5 warm-up epochs. The learning rate is decayed following a cosine scheduler down to 10^{-6}. Without other specification, we train the models for 300 epochs. Rand Augment [8] and Mixup [34] are both used by default. Following [25], we switch TokenMix and Mixup with the probability of 0.5. For training architectures with smaller model sizes, e.g., DeiT-T [25], PVT-T [30], or CaiT-XXS [26], we sample the λ in Equation 2 from a beta distribution Beta$(1.0, 1.0)$. We use binary cross-entropy (BCE) loss instead of the typical cross-entropy (CE) loss by default following [31, 2], as the mixed images are more likely to contain multiple labels. To generate the neural activation maps, we defautly use NFNet-F6 [3] following [14].

For transferring to the ADE20K dataset, we follow the setting in BEiT [1], and fine-tune for 160K steps with Adam [17] optimizer. The detailed hyperparameters are described in supplementary materials.

5 Main Results

5.1 ImageNet Results

We report the results on ImageNet-1K dataset with our TokenMix. As shown in Table 1, TokenMix consistently improves CutMix on various transformer-based architectures, i.e., DeiT [25], PVT [30], CaiT [26], and Swin Transformer [19]. Specifically, TokenMix outperforms CutMix [32] by +1% for DeiT, across DeiT-T to DeiT-B. We also improve popular hierarchical transformer architectures Swin-T and PVT-T for +0.4% and +0.5%, respectively. All the results demonstrate the effectiveness and generalization of the proposed TokenMix.

Our proposed TokenMix consists of two parts, i.e., token-level mixing and label refinement. We decouple the two parts and then compare them with the previous methods by fixing one part. In Table 2, we compare TokenMix to ReLabel [33] and TokenLabeling [14] with the same data augmentation method. The two methods utilize pixel-level supervision, but our TokenMix summarizes

Table 2: Comparisons with ReLabel and TokenLabeling with DeiT-T on ImageNet. GPU time refers to the increase of training time.

Augmentation	Supervision	Top-1 Acc.	GPU Time
CutMix	ImageNet	72.2	+0.0%
TokenMix	ImageNet	**72.7**	+0.0%
TokenMix	ReLabel	72.7	+0.8%
TokenMix	TokenLabeling	72.9	+0.8%
TokenMix	TokenMix	**73.2**	+0.8%

Table 3: Comparisons with previous mixing methods with DeiT-T on ImageNet.

Augmentation	Top-1 Acc.
CutMix [32]	72.2
Co-Mix [15]	72.2
SaliencyMix [27]	71.8
Puzzle-Mix [16]	72.3
TokenMix	**72.7**

Table 4: Transferring the pre-trained models to downstream semantic segmentation task on ADE20K dataset. TL and RL denote TokenLabeling and ReLabel respectively. ✓+RL/TL represents row 3/4 in Table 2.

Model	TokenMix	mIoU(%)	mAcc(%)	+ms mIoU(%)	+ms mAcc(%)
DeiT-T	✗	36.4	46.7	37.5	47.1
	✓+RL	36.6	47.0	38.1	47.9
	✓+TL	36.9	47.1	38.3	48.1
	✓	**37.1**	**47.5**	**38.6**	**48.2**
DeiT-S	✗	42.3	52.8	43.7	53.8
	✓	**44.5**	**55.0**	**45.9**	**56.1**
DeiT-B	✗	46.3	56.5	47.7	57.6
	✓	**46.8**	**56.9**	**48.2**	**58.1**

neural activations to create image-level target scores and is, therefore, more robust to individual pixel-level errors. Note that we use the same teacher network, i.e., NFNet-F6, to generate the offline targets. As shown in Table 2, TokenMix outperforms both ReLabel (+0.5%) and TokenLabeling (+0.3%) with the same training cost. We further compare TokenMix to previous mixing-based augmentation methods in Table 3. For a more fair comparison, we only use the labels from ImageNet. As shown in Table 3, TokenMix have performance advantages compared to other approaches. We see that the methods that introduce more foreground regions fail to improve CutMix on Vision Transformer. In contrast, our proposed TokenMix improves CutMix for +0.5% accuracy.

5.2 Transfer to Downstream Task

Pre-training on ImageNet-1K then finetuning to the downstream tasks is a common practice for many visual recognition tasks. It is important to verify whether the better pre-train with TokenMix can boost the performance on the downstream task. To do so, we transfer our TokenMix pre-trained models to the semantic segmentation task and compare them with regular pre-train. Note that TokenMix does not introduce extra computation overhead in the transfer stage. As shown in Table 4, we find that better pre-train from TokenMix consistently improves the segmentation performance on the ADE20K dataset. Notably, we

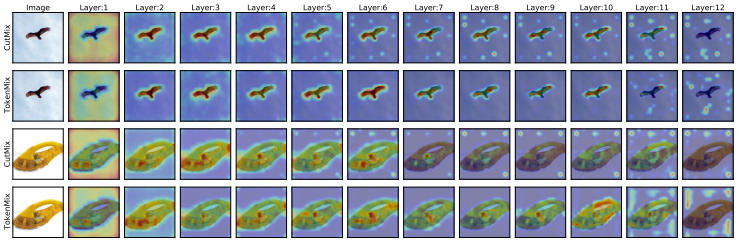


Fig. 3: Visualization of the attention maps of the class token in DeiT-S to attend to patch tokens at different layers. Using CutMix distracts the attention to background areas in the several middle layers. In contrast, the proposed TokenMix helps the class token focus more on foreground objects and leads to consistent performance gain.

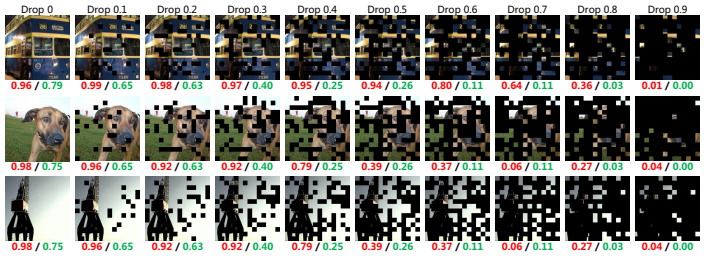


Fig. 4: Example images and the predicted confidences under different occlusion ratios. Red scores under the images are predicted by TokenMix, and green ones by CutMix. The model trained with TokenMix holds high confidence when a large number of patches are dropped, while the model trained with CutMix outputs low confidence.

improve DeiT-T for +0.7% mIoU, DeiT-S for +2.2% mIoU, DeiT-B for +0.5% mIoU. We notice that the performance gap becomes even larger (e.g. +1.1% mIoU for DeiT-T) when using multi-scale testing. All the results demonstrate the transferability of our TokenMix pre-trained models.

5.3 Main Properties

Besides the performance gains, we find that our proposed TokenMix improves transformers to be robust to occlusion, and focus more on the foreground area. All the visualization and analysis are conducted on DeiT-S.

TokenMix helps transformers focus on the foreground area. As discussed in Section 3, CutMix assigns targets of the mixed images based on linear combinations of labels of the pairs of mixing images, which might be inaccurate if the foreground region is cut. We find that the inaccurate labels make transformers pay incorrect attention to the input image. As shown in Figure 3, using CutMix distracts the transformer's attention to background areas in several middle

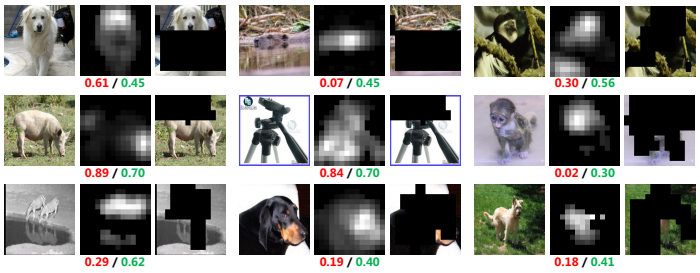


Fig. 6: The target scores generated by TokenMix and CutMix. For each tripled sub-figure, the left is the input image, the middle is the neural activation map, and the right is the masked image. Our approach generates more reasonable target scores, especially when the foreground region is cropped.

layers (layers 5-10). In comparison, TokenMix helps transformers learn to pay more attention to the foreground areas and leads to consistent performance gain.

TokenMix enhances the occlusion robustness of vision transformers. After training converges, we construct a sequence of images with different occlusion ratios. Specifically, we gradually drop 10% more patches and set the pixels inside to zero, and use the images for testing. We report the top-1 accuracy on ImageNet under different drop ratios. As shown in Figure 5, the model trained with TokenMix surpasses that with CutMix by increasingly larger margins as the drop ratio grows, demonstrating its better occlusion robustness. Specifically, we notice a ∼10% performance gap at the drop ratio of 80%. We further visualize some examples in Figure 4. It can be found that when about 40% tokens are dropped, the predicted ground-truth class confidences of the baseline model decrease to very low values (0.25 in the first row) while the model trained with our TokenMix holds higher confidences (0.95 in the first row).

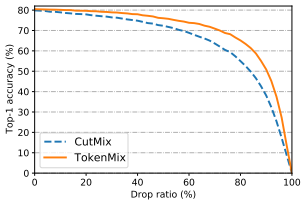


Fig. 5: ImageNet top-1 accuracy of DeiT-S under different drop ratios. The gap between the model trained with CutMix and that with TokenMix increases as the drop ratio grows.

6 Ablative studies

In this section, we conduct various ablation studies to analyze our proposed TokenMix. We use DeiT-S as the backbone and train it on ImageNet for 300 epochs unless otherwise specified. All other training settings are the same as described in Section 4. We report the top-1 accuracy on ImageNet.

Table 6: Comparison of different ways to generate neural activation maps. NFNet-F6 is used by default.

Teacher	Teacher Top-1 Acc.	Top-1 Acc.
NFNet-F6 [3]	86.1	80.8
ResNet101 [12]	82.3	80.7
ResNet26 [12]	79.8	80.5
Saliency [21]	N/A	80.1

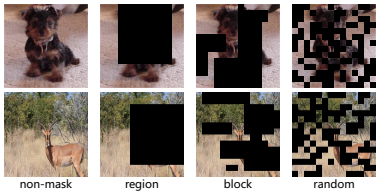
Table 7: Ablation of the way to generate target scores. Our approach generates target scores with content-based neural activation maps (denoted as *refinement*).

Model	Refinement	Top-1 Acc.
DeiT-S [25]	✗	79.8
	✓	**80.5** (+0.7)
Swin-T [19]	✗	81.2
	✓	**81.5** (+0.3)
ResNet50 [12]	✗	79.3
	✓	**79.8** (+0.5)

Table 8: Ablation of mask sampling strategy. The *region-based* strategy works best on ResNet50, but degrades on DeiT-S.

Model	region	random	block
DeiT-T [25]	72.2	**72.7**	**72.7**
DeiT-S [25]	79.8	**80.6**	**80.6**
ResNet50 [12]	79.3	78.3	**79.7**

Fig. 7: Illustration of different mask sampling strategies.



Integrating TokenMix with previous mixing-based methods. Table 5 presents the results of combining TokenMix with other mixing-based methods to train DeiT-B. When two mixing augmentations are utilized during training, one of them is randomly chosen to be used for data augmentation with a probability of 0.5 at each iteration. The baseline (row 1 in Table 5) does not use any mixing-based augmentations. Using only MixToken improves the baseline and CutMix by large margins. Specifically, TokenMix improves the baseline by +5.7% top-1 accuracy. Using both TokenMix and Mixup improves TokenMix-only by +1.4% top-1 accuracy. In contrast, using both CutMix and Mixup also improves top-1 accuracy to 81.8%, which, however, is still lower than TokenMix + Mixup.

Table 5: Performances of using a single or randomly sampled one of the multiple mixing methods for training DeiT-B.

Mixup [34]	CutMix [32]	TokenMix	Top-1 Acc.
✗	✗	✗	75.8
✗	✓	✗	78.7
✓	✗	✗	80.0
✗	✗	✓	**81.5**
✓	✓	✗	81.8
✗	✓	✓	82.0
✓	✗	✓	**82.9**

Different ways to generate neural activation maps. Table 6 presents the results of using different ways to generate neural activation maps. Besides our default choice of NFNet-F6, popular ResNet and hand-crafted saliency method are also compared. As shown in Table 6, TokenMix follows the general behavior of distillation techniques, i.e., if the performance of the teacher model is high, so is the student model. In addition, TokenMix is robust to different choices

of teacher networks for generating target scores. Even when the performance of the teacher drops by 6.3%, only 0.3% performance drop is observed on DeiT trained with our proposed TokenMix. However, using neural networks to generate neural activation maps is consistently better than using scores from hand-crafted approaches [21] as the targets generated by a teacher model are generally better to learn than hand-crafted ones. We further visualized the target scores of the mixed images in Figure 6. For each tripled sub-figure, the left is the input images, the middle is the neural activation maps, and the right is the masked images. The scores generated from CutMix are shown in green, while the red scores are generated by our approach. As shown in Figure 6, the target scores generated by our approach are more reasonable, especially when the foreground is cut.

Our neural-activation target scores are comptible with CutMix. To test whether our proposed target scores are compatible with CutMix, we conduct experiments of using CutMix to mix pairs of images but generating targets with our approach (denoted as *refinement* in Table 7), and train with various backbones, e.g., DeiT-S, Swin-T, and ResNet50. As shown in Table 7, we achieve consistent performance gains on those backbones. Specifically, we improve DeiT-S for +0.7%, Swin-T for +0.3%, and ResNet50 for +0.5% with nearly no extra computation cost during the training process. All the results verify the compatibility of our proposed target score assignment with CutMix.

Mask sampling strategy. Table 8 presents the impact of different sampling strategies on transformers and Convolution Neural Networks, as illustrated in Figure 7. The *region-based* sampling, widely used in [32, 21, 29], cuts a single large rectangular area from the mixing images. Our proposed TokenMix directly cuts at token-level. We compare two settings of our approach, masking multiple blocks (*block-based*) following our description in Section 3.2 or masking individual tokens separately (*random*). To better inspect the impact of sampling strategies alone, in all the experiments in Table 8, we directly use target scores generated by CutMix, instead of ours. As shown in Table 8, the *region-based* strategy achieves decent performance on ResNet50, but degrades on transformers, which validates our argument on the sub-optimality of using region-based cut for training transformers. Compared to the *random* strategy, the *block-based* strategy achieves similar performances on transformers, but performs much better on ResNet50. When using our proposed target score assignment approach, we further compare the *random* and the *block-based* sampling strategies. As shown in Table 9, the *block-based* strategy used in our final solution consistently has higher accuracy. The results further verify the effectiveness of our proposed TokenMix.

Table 9: Ablation of mask sampling strategy. The *block-based* strategy obtains higher accuracy with label refinement.

Model	Mask	Refinement	Top-1 Acc.
DeiT-T	random	✗	72.7
		✓	**72.9** (+0.2)
	block	✗	72.7
		✓	**73.2** (+0.5)
DeiT-S	random	✗	80.6
		✓	80.6
	block	✗	80.6
		✓	**80.8** (+0.2)

Table 10: Ablation of training epochs. TokenMix enjoys longer training. The extra 100 epochs of training improve +0.4% accuracy.

Mixing Method	Epoch	Top-1 Acc.
CutMix [32]	300	79.8
	400	79.9 (+0.1)
TokenMix	300	80.8
	400	**81.2** (+0.4)

Table 11: Ablation of the loss function. Binary cross-entropy (BCE) improves TokenMix, compared with multi-class cross-entropy (CE).

Mixing Method	Loss Type	Top-1 Acc.
CutMix [32]	CE	79.8
	BCE	79.8
TokenMix	CE	80.3
	BCE	**80.8** (+0.5)

Training epochs. Table 10 presents the results of longer training. As targets generated by a teacher network's neural activation maps can provide more appropriate scores and more challenging samples for training the transformers, which mitigates the risk of over-fitting scheme, our proposed TokenMix can enjoy longer training. As shown in Table 10, our TokenMix improves DeiT-S for +0.4% with additional 100 training epochs, while using CutMix for long training is less beneficial.

Loss function. As the mixed images may contain multiple objects of different classes, we adopt the binary cross-entropy (BCE) loss instead of the typical cross-entropy (CE) loss [31, 2]. Using BCE loss improves DeiT-S for +0.5% accuracy (Table 11) when training with our proposed TokenMix, as the cut-and-paste operation might generate a mixed image with multiple objects of different classes. It might be because the generated targets by CutMix are sub-optimal, we do not notice performance improvement of replacing CE with BCE when training DeiT-S with CutMix augmentation.

7 Conclusions

In this paper, we propose TokenMix, a token-level augmentation strategy that generalizes well across various transformer-based architectures. TokenMix is motivated by two key observations: 1) region-level mixing is less beneficial for transformer-based architectures, and 2) assigning target of mixed images with linear combination might be inaccurate and even counter-intuitive. Our proposed TokenMix directly cuts at token level and obtains the target of the mixed images with content-based neural activation maps. Empirical results show that TokenMix has the properties of enhancing occlusion robustness and helping vision transformers focus on the foreground area of input images. Besides, TokenMix consistently improves various transformer-based architectures, including DeiT, PVT, and Swin Transformer.

Acknowledgement Hongsheng Li is also a Principal Investigator of Centre for Perceptual and Interactive Intelligence Limited (CPII). This work is supported in part by CPII, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants (Nos. 14204021, 14207319), in part by CUHK Strategic Fund.

References

1. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
2. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., Oord, A.v.d.: Are we done with imagenet? arXiv preprint arXiv:2006.07159 (2020)
3. Brock, A., De, S., Smith, S.L., Simonyan, K.: High-performance large-scale image recognition without normalization. arXiv preprint arXiv:2102.06171 (2021)
4. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 782–791 (2021)
5. Chen, J., Sun, S., He, J., Torr, P.H.S., Yuille, A.L., Bai, S.: Transmix: Attend to mix for vision transformers. ArXiv **abs/2111.09833** (2021)
6. Chen, P., Liu, S., Zhao, H., Jia, J.: Gridmask data augmentation. ArXiv **abs/2001.04086** (2020)
7. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)
8. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703 (2020)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Devries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. ArXiv **abs/1708.04552** (2017)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. ArXiv **abs/1912.02781** (2020)
14. Jiang, Z., Hou, Q., Yuan, L., Zhou, D., Jin, X., Wang, A., Feng, J.: Token labeling: Training a 85.5% top-1 accuracy vision transformer with 56m parameters on imagenet. arXiv preprint arXiv:2104.10858 (2021)
15. Kim, J.H., Choo, W., Jeong, H., Song, H.O.: Co-mixup: Saliency guided joint mixup with supermodular diversity. arXiv preprint arXiv:2102.03065 (2021)
16. Kim, J.H., Choo, W., Song, H.O.: Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In: International Conference on Machine Learning. pp. 5275–5285. PMLR (2020)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)

19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Montabone, S., Soto, A.: Human detection using a mobile platform and novel features derived from a visual saliency mechanism. Image and Vision Computing **28**(3), 391–402 (2010)
22. Qin, J., Fang, J., Zhang, Q., Liu, W., gang Wang, X., Wang, X.: Resizemix: Mixing data with preserved object information and true labels. ArXiv **abs/2012.11101** (2020)
23. Singh, K.K., Yu, H., Sarmasi, A., Pradeep, G., Lee, Y.J.: Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond. ArXiv **abs/1811.02545** (2018)
24. Takahashi, R., Matsubara, T., Uehara, K.: Data augmentation using random image cropping and patching for deep cnns. IEEE Transactions on Circuits and Systems for Video Technology **30**, 2917–2931 (2020)
25. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
26. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. arXiv preprint arXiv:2103.17239 (2021)
27. Uddin, A., Monira, M., Shin, W., Chung, T., Bae, S.H., et al.: Saliencymix: A saliency guided data augmentation strategy for better regularization. arXiv preprint arXiv:2006.01791 (2020)
28. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. In: ICML (2019)
29. Walawalkar, D., Shen, Z., Liu, Z., Savvides, M.: Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. arXiv preprint arXiv:2003.13048 (2020)
30. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv preprint arXiv:2102.12122 (2021)
31. Wightman, R., Touvron, H., Jégou, H.: Resnet strikes back: An improved training procedure in timm. arXiv preprint arXiv:2110.00476 (2021)
32. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032 (2019)
33. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-labeling imagenet: from single to multi-labels, from global to localized labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2340–2350 (2021)
34. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
35. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. ArXiv **abs/1708.04896** (2020)
36. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)

37. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)