**TransMix: Attend to Mix for Vision Transformers**

Jie-Neng Chen\(^1\)\(^*\) Shuyang Sun\(^2\)\(^*\) Ju He\(^1\) Philip Torr\(^2\) Alan Yuille\(^1\) Song Bai\(^3\)

\(^1\)Johns Hopkins University\(^1\)\(^2\)University of Oxford\(^3\)ByteDance Inc.

**Abstract**

Mixup-based augmentation has been found to be effective for generalizing models during training, especially for Vision Transformers (ViTs) since they can easily overfit. However, previous mixup-based methods have an underlying prior knowledge that the linearly interpolated ratio of targets should be kept the same as the ratio proposed in input interpolation. This may lead to a strange phenomenon that sometimes there is no valid object in the mixed image due to the random process in augmentation but there is still response in the label space. To bridge such gap between the input and label spaces, we propose TransMix, which mixes labels based on the attention maps of Vision Transformers. The confidence of the label will be larger if the corresponding input image is weighted higher by the attention map. TransMix is embarrassingly simple and can be implemented in just a few lines of code without introducing any extra parameters and FLOPs to ViT-based models. Experimental results show that our method can consistently improve various ViT-based models at scales on ImageNet classification. After pre-trained with TransMix on ImageNet, the ViT-based models also demonstrate better transferability to semantic segmentation, object detection and instance segmentation. TransMix also exhibits to be more robust when evaluating on 4 different benchmarks. Code is publicly available at https://github.com/Beckschen/TransMix.

**1. Introduction**

Transformers [42] have been dominant in nearly all tasks in natural language processing. Recently, transformer-based architectures like Vision Transformer (ViT) [12] have been introduced into the field of computer vision and show great promise on tasks like image classification [12, 13, 30, 40], object detection [48, 30, 15] and image segmentation [48, 30, 37]. However, recent works have found that ViT-based networks are hard to optimize and can easily overfit if the training data is not sufficient. A quick solution [48, 30, 37]. However, recent works have found that sometimes there is no valid object in the mixed image due to the random process in augmentation but there is still response in the label space. To bridge such gap between the input and label spaces, we propose TransMix, which mixes labels based on the attention maps of Vision Transformers. The confidence of the label will be larger if the corresponding input image is weighted higher by the attention map. TransMix is embarrassingly simple and can be implemented in just a few lines of code without introducing any extra parameters and FLOPs to ViT-based models. Experimental results show that our method can consistently improve various ViT-based models at scales on ImageNet classification. After pre-trained with TransMix on ImageNet, the ViT-based models also demonstrate better transferability to semantic segmentation, object detection and instance segmentation. TransMix also exhibits to be more robust when evaluating on 4 different benchmarks. Code is publicly available at https://github.com/Beckschen/TransMix.

As shown in Figure 1, pixels in the background will not contribute to the label space as equally as those in the salient area. Some existing works [45, 41, 28] also find this problem and solve it by means of only mixing the most descriptive parts on the input level. Nevertheless, manipulating on inputs with the above methods may narrow the space of augmentation since they tend to less consider to put the background image into the mixture. Meanwhile, the above methods cost more number of parameters and/or training throughput to extract the salient region of input. For example, Puzzle-Mix [28] requires model to forward and backward twice in an iteration and Attentive-Cutmix [45] introduce a 24M external CNN to extract salient features.

Instead of investigating how to better mix images on the input level, in this paper, we focus more on how to mild the...
We show that such frustratingly simple idea can lead to consistent and remarkable improvement for a wide range of tasks and models. As exhibited in Figure 2, TransMix can steadily boost all the listed ViT-based models. Notably, TransMix can further lift the top-1 accuracy on ImageNet by 0.9% for both DeiT-S and a large variant XCiT-L. Interestingly, the largest model XCiT-L gains the most among all XCiT model scales.

Moreover, we demonstrate that if the model is first pre-trained with TransMix on ImageNet, the superiorit can be further transferred onto downstream tasks including object detection, instance segmentation, semantic segmentation, and weakly-supervised object segmentation/localization. We also observe that TransMix can help the model to be more robust after evaluating it on 4 different benchmarks.

2. Related Work

Vision Transformers (ViTs). Recently, Vision Transformer (ViT) [12] was proposed to adapt the Transformer for image recognition by tokenizing and flattening images into a sequence of tokens. ViT is based on a sequence of Transformer blocks consisting of multi-head self-attention layers and feed-forward networks. DeiT [39] strengthens ViT by introducing a powerful training recipe and adopting knowledge distillation. Built upon the success of ViT, many efforts have been devoted to improving ViT and adapting it into various vision tasks including image classification [39, 40, 13, 50, 24, 30, 18], object localization/detection [16, 48, 30, 15] and image segmentation [48, 30, 37, 5].

Mixup and its variants. Data augmentation has been widely studied to prevent DeepNets from over-fitting to the training data. To train and improve vision Transformer stably, Mixup and CutMix are two of the most helpful augmentation methods [39]. Mixup [54] is a successful image mixture technique that obtains an augmented image by pixel-wisely weighted combination of two global images. The following Mixup variants [44, 36, 17, 22, 53, 45, 41, 28] can be categorized into global image mixture (e.g. Manifold-Mixup [44], Un-Mix [36]) and regional image mixture (e.g. CutMix [53], Puzzle-Mix [28], Attentive-CutMix [45] and SaliencyMix [41]). Among all Mixup variants, the saliency-based methods including the attentive-CutMix, puzzle-Mix and saliency-CutMix are the most similar ones to our approach. However, TransMix has two fundamental differences with them: (1) Previous saliency-based methods e.g. [28, 45, 41] enforce the image patch cropped in a salient region of the input image. Instead of manipulating in the input space, our TransMix focuses on how to more accurately assigning labels in the label space. (2) Previous saliency-based methods like [45] may use extra parameters to extract the saliency region. TransMix naturally exploits the Transformer’s attention mechanism without any extra parameters. Experimental results also show that TransMix can lead to better results on ImageNet compared with these methods.

Data-adaptive loss weight assignment. TransMix reassigns the ground truth labels with attentional guidance, which is related to data-adaptive loss weight assignment. Some existing works have found that the attention-like information can help to alleviate the long-tail problems for tasks like point cloud analyzing [31], instance segmentation [47], image demosaicing [38] etc.

3. TransMix

3.1. Setup and Background

CutMix data augmentation. CutMix is a simple data augmentation technique combining two input-label pairs
\((x_A, y_A)\) and \((x_B, y_B)\) to augment a new training sample \((\tilde{x}, \tilde{y})\). Formu

\[
\tilde{x} = M \odot x_A + (1 - M) \odot x_B, 
\]

\[
\tilde{y} = \lambda y_A + (1 - \lambda) y_B,
\]

where \(M \in \{0, 1\}^{HW}\) denotes a binary mask indicating where to drop out and fill in from two images, \(1\) is a binary mask filled with ones, and \(\odot\) is element-wise multiplication. \(\lambda\) is the proportion of \(y_A\) in the mixed label.

During augmentation, a randomly sampled region in \(x_B\) is removed and filled in with the patch cropped from \(A\) of \(x_A\), where the patch’s bounding box coordinates are uniformly sampled as \((r_x, r_y, r_w, r_h)\). The mixed-target assignment factor \(\lambda\) is equal to the cropped area ratio \(\frac{r_w r_h}{WH}\).

**Self-attention** Self-attention, as introduced by [43], operates on an input matrix \(x \in \mathbb{R}^{N \times d}\), where \(N\) is the number of tokens, each of dimensionality \(d\). The input \(x\) is linearly projected to queries, keys, and values, using the weight matrices \(w_q \in \mathbb{R}^{d \times d_q}, w_k \in \mathbb{R}^{d \times d_k}\) and \(w_v \in \mathbb{R}^{d \times d_v}\), such that \(q = xw_q, k = xw_k\) and \(v = xw_v\), where \(d_q = d_k\). Queries and keys are used to compute an attention map \(A(q, k) = \text{Softmax}(qk^\top / \sqrt{d}) \in \mathbb{R}^{N \times N}\), and the output of the self-attention operation is defined as the weighted sum of \(N\) token features in \(v\) with the weights corresponding to the attention map: 

\[
\text{Attention}(q, k, v) = A(q, k)v.
\]

Single-head self-attention can be extended to multi-head self-attention by linearly projecting the queries, keys and values \(g\) times with different, learned linear projections to \(d_k, d_k, d_v\) dimensions, respectively.

### 3.2. TransMix

We propose TransMix to assign mixup labels with the guidance of attention map, where the attention map is defined specifically as the **multi-head class attention** \(A\), which is calculated as a part of self-attention. In the classification task, a class token is a query whose corresponding keys are the all input tokens, and class attention \(A\) is the attention map from the class token to the input tokens, summarizing which input tokens are the most useful to the final classifier. We then propose to use the class attention \(A\) to mix labels.

**Multi-head Class Attention** Vision Transformers (ViTs)\[12\] divide and embed an image \(x \in \mathbb{R}^{3 \times H \times W}\) to \(p\) patch tokens \(x_{\text{patches}} \in \mathbb{R}^{p \times d}\), and aggregate the global information by a class token \(x_{\text{cls}} \in \mathbb{R}^{1 \times d}\), where \(d\) is the dimension of embedding. ViTs operate on the patch embedding \(z = [x_{\text{cls}}, x_{\text{patches}}] \in \mathbb{R}^{(1+p) \times d}\).

Given a Transformer with \(g\) attention heads and input patch embedding \(z\), we parametrize the multi-head class-attention with projection matrices \(w_q, w_k \in \mathbb{R}^{d \times d}\). The class attention for each head can be formulated as:

\[
q = x_{\text{cls}} \cdot w_q,
\]

\[
k = z \cdot w_k,
\]

\[
A' = \text{Softmax}(q \cdot k^\top / \sqrt{d/g}),
\]

\[
A = \{A_{i}^{\prime} | i \in [1, p]\},
\]

where \(q \cdot k^\top \in \mathbb{R}^{1 \times (1+p)}\) indicates the class token is a query whose corresponding keys are the all input tokens, and \(A \in [0, 1]^p\) is the attention map from the class token to the image patch tokens, summarizing which patches are the most useful to the final classifier. When there are multiple heads in the attention, we simply average across all attention heads to obtain \(A \in [0, 1]^p\). In implementation, \(A\) in Eqn. \((6)\) is available as an intermediate output from the last Transformer block without architecture modification.

**Mixing labels with the attention map** \(A\) We follow the process of input mixture proposed in CutMix, which is defined in Eqn. \((1)\), then we re-calculate \(\lambda\) (the proportion of \(y_A\) in Eqn. \((2)\)) with the guidance of the attention map \(A\):

\[
\lambda = A \cdot \downarrow (M).
\]

Here \(\downarrow (\cdot)\) denotes the nearest-neighbor interpolation down-sampling that can transform the original \(M\) from \(HW\) into \(p\) pixels. Note that we omit the dimension unsqueezing in Eqn. \((7)\) for simplicity. In this way, the network can learn to re-assign the weight of labels for each data point dynamically based on their responses in the attention map. The input that is better focused by the attention map will be assigned with a higher value in the mixed label.

### 3.3. Pseudo-code

Algorithm 1 provides the pseudo-code of TransMix in a pytorch-like style. The clean pseudo-code shows that simply few lines of code can boost the performance in the plug-and-play manner.

### 4. Experiments

In this section, we mainly demonstrate the effectiveness, transferability, robustness, and generalizability of TransMix. We verify the effectiveness of TransMix on ImageNet-1k classification in Section 4.1 and the transferability onto downstream tasks including semantic segmentation, object detection, and instance segmentation in Section 4.2. The robustness of TransMix is examined on 4 benchmarks in Section 4.3. Interestingly, we discover the mutual effects of TransMix and attention in Section 4.4. We validate the generalizability to Swin Transformer which is lacking class-token in Section 4.5. Lastly, TransMix is compared with the state-of-the-art Mixup augmentation variants in Section 4.6.
Implementation Details

We use ImageNet-1k [11] to train and evaluate our methods for image classification. ImageNet-1k consists of 1.28M training images and 50k validation images, labeled across 1000 semantic categories. The implementation is based on the Timm [49] library. Unless specified otherwise, we make minimal changes to the official papers’ implementations.

All Transformers are trained for 300 epochs expect that [author?] [13] and [author?] [40] report 400 epochs for Xcit and CaiT respectively. As deploying DeiT [39] training scheme, all baselines have already contained the carefully tuned regularization methods including RandAug [10], Stochastic Depth [26], Mixup [54] and CutMix [53]. To ease implementation, TransMix shares the same cropped region with CutMix for the input, whereas the label assignment is the mean of both methods. We throw away repeated augmented [25] due to its negative effects examined in [24]. We set warmup epoch to 20 expect DeiT-B keeping 5. The accuracy of our baseline implementation fluctuates only by ±0.1% compared with results reported in DeiT [39].

The attention map A in Eqn. 6 can be obtained as an intermediate output from the multi-head self-attention layer of the last Transformer block.

Results

As shown in Table 1, TransMix can steadily boost the top-1 accuracy on ImageNet for all the listed models. No matter how complex the model is, TransMix can always help to boost the baseline performance. Note that these models are with a wide range of model complexities, and the baselines are all carefully tuned with various data augmentation techniques e.g. RandAug [10], Mixup [54] and CutMix [53]. To be specific, TransMix can promote the top-1 accuracy of the small variant DeiT-S by 0.9%. Benefit from the higher attention quality, TransMix can also lift the top-1 accuracy of the large model Xcit-L by a remarkable 0.9%. We emphasize that these systematic improvements with just a tiny tweak on data augmentation is significant when compared with the structural modification on models. For example, CrossViT-B [4] only lifts the DeiT-B baseline result by 0.4% with 20.9% parameters overhead while TransMix leads to more improvement in a parameter-free style. Particularly, TransMix consistently boosts the base/large variants in the range of 0.6% to 0.9%, which is more striking than engineering new architectures such as PiT-B [24], T2T-24 [51], CrossViT-B [4] with the gains of 0.2%, 0.5%, 0.4% respectively.

4.2. Transfer to Downstream Tasks

ImageNet pre-training is the de-facto standard practice for many visual recognition tasks [19]. Before training for downstream tasks, the weights pre-trained on ImageNet is used to initialize the Transformer backbone. We demonstrate the transferability of our TransMix-based pre-trained models on the downstream task including semantic segmentation, object detection and instance segmentation, on which we observe the improvements over the vanilla pre-trained baselines.

Semantic Segmentation

In our experiments, the sequence of patch encoding $z_{\text{patches}} \in \mathbb{R}^{P \times d}$ is decoded to a segmentation map $s \in \mathbb{R}^{H \times W \times K}$ where K is the number of semantic classes. We adopt two convolution-free de-
The reason for adopting the Linear decoder is to preserve the pre-trained information to the greatest extent. For linear decoder, a point-wise linear layer on DeiT patch encoding \( z_{\text{patches}} \in \mathbb{R}^{p \times d} \) is used to produce patch-level logits \( z_{\text{lin}} \in \mathbb{R}^{p \times K} \), which are reshaped and bilinearly upsampled to segmentation map \( s \). The Segmenter [37] decoder is a Transformer-based decoder namely Mask Transformer introduced in [37, 46].

We train and evaluate the models on the Pascal Context [33] dataset and report Intersection over Union (mIoU) averaged over all classes as the main metric. The training set contains 4998 images with 59 semantic classes plus a background class. The validation set contains 5105 images. The training scheme follows [33] which is built on MM-SEGmentation [9]. As a reference, the result of ResNet101-Deeplabv3+ [7, 8] is reported in MMSegmentation [9].

According to Table 2, TransMix pre-trained DeiT-S-Linear and DeiT-S-Segmenter improve over the vanilla pre-trained baselines by 0.6% and 0.9% mIoU respectively. There are consistent improvements on multi-scale testing.

### Object Detection and Instance Segmentation

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Decoder</th>
<th>TransMix-Pretrained</th>
<th>mAcc</th>
<th>mIoU</th>
<th>mIoU (MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet101 [21]</td>
<td>Deeplabv3+ [8]</td>
<td>57.4</td>
<td>47.3</td>
<td>48.5</td>
<td></td>
</tr>
<tr>
<td>DeiT-S [39]</td>
<td>Linear</td>
<td>59.4</td>
<td>49.1</td>
<td>49.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Segmenter [37]</td>
<td>✓</td>
<td>60.2</td>
<td>49.7</td>
<td>50.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>60.4</td>
<td>49.7</td>
<td>50.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Overhead-free impact of TransMix on transferring to downstream semantic segmentation task on the Pascal Context [33] dataset. (MS) denotes multi-scale testing.

### 4.3. Robustness Analysis

Recently, the discussions regarding the robustness of vision Transformer are emerging [34, 32, 1]. To verify if TransMix can improve ViT-based models’ robustness and out-of-distribution performance, we evaluated our TransMix pre-trained models on four robustness scenarios including occlusion, spatial structure shuffling, natural adversarial example, and out-of-distribution detection.

#### Robustness to Occlusion (author?)

[34] studies whether ViTs perform robustly in occluded scenarios, where some or most of the image content is missing. To be specific, vision Transformers divide an image into \( M=196 \) patches belonging to a \( 14 \times 14 \) spatial grid; i.e. an image of size \( 224 \times 224 \times 3 \) is split into 196 patches, each of size \( 16 \times 16 \). Patch Dropping means replacing original image patches with blank 0-value patches. As an example, dropping 100 such patches from the input is equivalent to losing 51% of the image content. Following [34], we showcase the classification accuracy on ImageNet-1k validation set with three dropping settings. (1) Random Patch Dropping: A subset of \( M \) patches is randomly selected and dropped. (2) Salient (foreground) Patch Dropping: This studies the robustness of ViTs against occlusions of highly salient regions. (author?) [34] thresholds DINO’s attention map to obtain salient patches, which are dropped by ratios. (3) Non-salient (background) Patch Dropping: The least salient regions of an image are selected and dropped following the same approach as above.

As shown in Figure 3, DeiT-S with TransMix outperform vanilla DeiT-S on all occlusion levels especially for extreme occlusion (information loss ratio >0.7).

#### Sensitivity to Spatial Structure Shuffling

We study the model’s sensitivity to the spatial structure by shuffling on
input image patches. Specifically, we randomly shuffle the image patches with different grid sizes following [34]. Note that a shuffle grid size of 1 means no shuffle, and a shuffle grid size of 196 means all patch tokens are shuffled. Figure 4 shows the consistent improvements over baseline, and the accuracy averaged on all shuffled grid sizes for TransMix-DeiT-S and DeiT-S are 62.8% and 58.4% respectively. The superior 4.2% gain indicates that TransMix enables Transformers rely less on positional embedding to preserve the most informative context for classification.

**Natural Adversarial Example** The ImageNet-A dataset [23] adversarially collects 7500 unmodified, natural but “hard” real-world images, which are drawn from some challenging scenarios (e.g., fog scene and occlusion). The metric for assessing classifiers’ robustness to adversarially filtered examples includes the top-1 accuracy, Calibration Error (CalibError) [29, 23], and Area Under the Response Rate Accuracy Curve (AURRA). CalibError judges how classifiers can reliably forecast their accuracy. AURRA is an uncertainty estimation metric introduced in [23]. As shown in Table 4, TransMix-trained DeiT-S is superior to vanilla DeiT-S on all metrics.

**Out-of-distribution Detection** The ImageNet-O [23] is an adversarial out-of-distribution detection dataset, which adversarially collects 2000 images from outside ImageNet-1k. The anomalies of unforeseen classes should result in low-confidence predictions. The metric is the area under the precision-recall curve (AUPR) [23]. Table 4 indicates that TransMix-trained DeiT-S outperform DeiT-S by 1% AUPR.

### 4.4. Mutual Effect of TransMix and Attention

**Will TransMix Benefit Attention?** To evaluate the quality of attention matrix, we directly threshold the class-token attention $A$ from DeiT-S to obtain a binary attention mask (the same with [3, 34]) with threshold 0.9 and then conduct two tasks including (1) **Weakly Supervised Automatic Segmentation** on Pascal VOC 2012 benchmark [14]. (2) **Weakly Supervised Object Localization (WOSL)** on ImageNet-1k validation set [35] where the bounding box annotations are only available for evaluation. For task (1), we compute the Jaccard similarity between ground truth and binary attention masks over the PASCAL-VOC12 validation set. For task (2), different from CAM-based methods for CNNs, we directly generate one tight bounding box from the binary attention masks, which is compared with ground-truth bounding box on ImageNet-1k. Both tasks are weakly-supervised since only the class-level ImageNet labels are used for training models (i.e. neither bounding box supervision for object localization nor per-pixel supervision for segmentation). The attention masks generated from TransMix-DeiT-S or vanilla DeiT-S are compared with ground-truth on these two benchmarks. The evaluated scores can quantitatively help us to understand if TransMix has a positive effect on the quality of attention map.

**Can Better Attention Nurture TransMix?** The experiments above prove that TransMix can benefit attention map, and it’s natural to ask that can better attention map nurture TransMix in return? We hypothesize that the better attention map is used, the more accurate TransMix adjusts the mixed-target assignment. For example, Dino [3] confirm that the attention maps obtained from the model via self-supervised training [3, 2] retain greater quality. To validate if a better attention map helps TransMix, we design an experiment that replaces the attention map with that generated from a parameter-frozen external model. The external parameter-frozen model can be (1) Dino self-supervised pre-trained DeiT-S (2) DeiT-S that is fully-supervised trained on ImageNet-1k. (3) DeiT-S that is fully-supervised trained with a knowledge distillation setting on ImageNet-1k. However, the results shown in Table 6 contradict the hypothesis.

**Intriguing Dynamic Property** With pre-trained Dino as the attention provider, the performance is slightly worse than that of self-serving. Training with attention guidance from a external fully-supervised parameter-frozen DeiT-S, TransMix suffers from a significant drop from 80.7% to 80.4% top-1 accuracy, though it is still better than vanilla model’s 79.8%. This phenomenon can ascribe to the dynamic property of TransMix, meaning that the per-iteration parameter update will dynamically diversify the self-attention for the same input image. In contrast, the parameter-frozen external models statically produce the same self-attention for an image, and thus undermine the regularization capability.

### Table 4. Model’s robustness against natural adversarial examples on ImageNet-A and out-of-distribution examples on ImageNet-O.

<table>
<thead>
<tr>
<th>Models</th>
<th>Top1-Acc</th>
<th>Calib-Error ↓</th>
<th>AURRA</th>
<th>AUPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeiT-S</td>
<td>19.1</td>
<td>32.0</td>
<td>23.8</td>
<td>20.9</td>
</tr>
<tr>
<td>TransMix-DeiT-S</td>
<td>21.1</td>
<td>31.2</td>
<td>28.8</td>
<td>21.9</td>
</tr>
</tbody>
</table>

### Table 5. Quantitative evaluation of the attention map. Segmentation JI denotes the Jaccard index for weakly supervised segmentation on Pascal VOC and Localization mIoU denotes the bounding box mIoU for weakly supervised object localization on ImageNet-1k.

<table>
<thead>
<tr>
<th>Models</th>
<th>Segmentation JI (%)</th>
<th>Localization mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeiT-S</td>
<td>29.2</td>
<td>34.9</td>
</tr>
<tr>
<td>TransMix-DeiT-S</td>
<td>29.9</td>
<td>44.4</td>
</tr>
</tbody>
</table>

12140
Figure 3. **Robustness against occlusion.** Model's robustness against occlusion with different information loss ratios is studied. 3 patch dropping settings: Random Patch Dropping (left), Salient Patch Dropping (middle), and Non-Salient Patch Dropping (right) are considered.

Figure 4. **Robustness against shuffle.** Model’s robustness against shuffle with different grid shuffle sizes is studied. (Placeholder)

<table>
<thead>
<tr>
<th>Models</th>
<th>Params</th>
<th>FLOPs</th>
<th>top-1 Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA-Swin-T [30, 40]</td>
<td>28.3M</td>
<td>4.2G</td>
<td>81.6</td>
</tr>
<tr>
<td>TransMix-CA-Swin-T</td>
<td>28.3M</td>
<td>4.2G</td>
<td>81.8</td>
</tr>
<tr>
<td>Swin-S [30]</td>
<td>49.6M</td>
<td>8.8G</td>
<td>83.0</td>
</tr>
<tr>
<td>CA-Swin-S [30, 40]</td>
<td>49.6M</td>
<td>8.5G</td>
<td>82.8</td>
</tr>
<tr>
<td>TransMix-CA-Swin-S</td>
<td>49.6M</td>
<td>8.5G</td>
<td>83.2</td>
</tr>
</tbody>
</table>

Table 7. Generalization to Swin Transformer [30] which lacks the class-token. CA denote the class attention block [40]. CA-Swin replaces Swin's last block with a CA block with fewer FLOPs.

### 4.5. Generalizability Study

One might be wondering if TransMix can be generalized to those models without the class token such as Swin-Transformer (Swin) [30]. Such models directly apply average pooling onto patch tokens to obtain logits, and therefore how much each patch token contributes to the final prediction is a black-box procedure without class attention $A$.

To tackle the aforementioned issue, we develop a Swin variant named as CA-Swin that replaces the last Swin block with a classification attention (CA) block without parameter overhead, which makes it possible to generalize TransMix onto Swin. Inspired by CaiT [40], the classification attention block aims at inserting the class token in a plug-and-play manner to those Transformers originally with only patch tokens, and make the classification attention $A$ accessible. We then compare the Swin-T, CA-Swin-T, TransMix-CA-Swin-T on ImageNet-1k with the same experimental setup in Sec. 4.1. All three models are at the same 28.3M parameters. TransMix-CA-Swin-T and CA-Swin-T have 7% fewer FLOPs than the baseline Swin-T. The top1 validation accuracy are 81.3%, 81.6% and 81.8% for Swin-T, CA-Swin-T and TransMix-CA-Swin-T, respectively. TransMix on Swin-S improves performance with fewer FLOPs as well. This preliminary study empirically proves the generalizability of TransMix.

### 4.6. Comparison with State-of-the-art Mixup Variants

In this section, we provide the comprehensive comparison with many state-of-the-art mixup variants on ImageNet-1k. This is the first time that compare these variants on vision Transformer in a fair setting. The implementation details for Mixup variants on top of DeiT-S are pro-
<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone Params</th>
<th>Speed (im/sec)</th>
<th>top-1 Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>22M</td>
<td>322</td>
<td>78.6</td>
</tr>
<tr>
<td>CutMix [53]</td>
<td>22M</td>
<td>322</td>
<td>79.8 (+1.2)</td>
</tr>
<tr>
<td>Attentive-CutMix [45]</td>
<td>DeiT-S</td>
<td>46M</td>
<td>239</td>
</tr>
<tr>
<td>SaliencyMix [41]</td>
<td>22M</td>
<td>314</td>
<td>79.2 (+0.6)</td>
</tr>
<tr>
<td>Puzzle-Mix [28]</td>
<td>22M</td>
<td>139</td>
<td>79.8 (+1.2)</td>
</tr>
<tr>
<td>TransMix</td>
<td>22M</td>
<td>322</td>
<td>80.7 (+2.1)</td>
</tr>
</tbody>
</table>

Table 8. Top1-accuracy, training speed (im/sec) and number of parameters comparison with state-of-the-art Mixup variants on ImageNet-1k. All listed models are built upon DeiT training recipe for fair comparison. Training speed (im/sec) takes account of data mixup, model forward and backward in train-time.

Table 8 shows TransMix significantly outperforms all other Mixup variants. The saliency-based methods (e.g. SaliencyMix and Puzzle-Mix) reveal no advantages to vision Transformer, compared to the vanilla CutMix. We analyze that these methods are cumbersomely tuned and face difficulty in transferring to new architecture. For example, Attentive-CutMix bring not only extra time but also parameter overhead as it introduces an external model to extract saliency map. Puzzle-Mix performs the lowest speed as it forward and backward twice during one training iteration. By contrast, TransMix yields a striking 2.1% performance advancement with the highest training throughput and no parameter-overhead.

**Ablation Study** Unlike suprisingly 8 hyper-parameters in PuzzleMix, our proposed TransMix exists very clean and introduces almost no hyper-parameter. Still, we conduct ablation study for TransMix regarding the attention map generation in the supplementary material, which shows that the default is the best.

**Visualization** We provide the visualization of TransMix as shown in Figure 5. For instance, the first row illustrate that the old area-based label assignment is counter-intuitive as image A's foreground is occluded by image B's patch, and TransMix corrects the label assignment via Transformer attention. TransMix is able to lift the label weight if the discriminative fine-grained attribute appears (e.g. Pomeranian dog's cheek and eyes in the second row).

**5. Conclusion**

In this paper, we present TransMix, a simple yet effective data augmentation technique that assigns Mixup labels with attentional guidance for Vision Transformers. TransMix naturally exploits Transformer’s attention map to assign the confidence for the mixed-target, and lifts the top-1 accuracy on ImageNet by 0.9% for both DeiT-S and a large variant XCiT-L. Extensive experiments are conducted to verify the effectiveness, transferability, robustness and generalizability of TransMix on totally 10 benchmarks.

**Limitations** Since we are the first work that pushes an extra mile for the Mixup-based methods towards augmenting vision Transformers, we indeed have limitations as follows: (1) TransMix can not handle well with those backbones without class token, as it strongly relies on the class attention. This limitation can be mitigated in Section 4.5 at the cost of architecture modification. (2) TransMix requires the attention map to be spatially aligned with the input, resulting in poor compatibility with deformable-based Transformer (e.g. PS-ViT [52], DeformDETR [55]). This can be potentially solved by calibrating attention map to the input spatial location by leveraging deformed offset grid.

**Acknowledgement** We would like to thank Xiaoyu Yue, Huiyu Wang, Qihang Yu and Chen Wei for their feedbacks. This work is done while the first two authors intern at Bytedance Inc. Jie-Neng Chen and Alan Yuille are supported by ONR N00014-21-1-2812, and the Lustgarten Foundation for Pancreatic Cancer Research. Shuyang Sun and Philip Torr are supported by the Turing AI Fellowship EP/W002981/1, ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1 and EPSRC/MURI grant EP/N019474/1. We would also like to thank the Royal Academy of Engineering and FiveAI.
References

[32] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. CVPR, 2022. 5
[38] 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 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 1, 2, 4, 5
[50] 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 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 6023–6032, 2019. 1, 2, 4, 8