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 [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 to this problem is to apply data augmentation and regularization techniques during training. Among them, the mixup-based methods like Mixup [54] and CutMix [53] are proven to be particularly helpful for generalizing the ViT-based network [39].

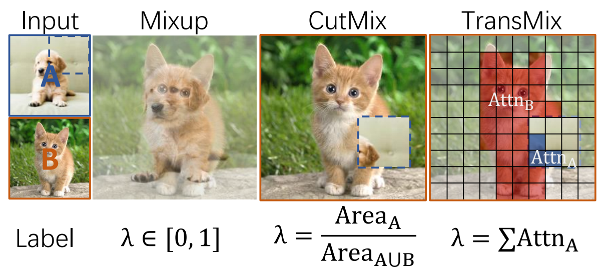

Figure 1. Mixup [54] and CutMix [53] samples λ (proportion of label $\mathbf{y}_A$) randomly from a Beta distribution, while our TransMix calculates λ with the sum of the values within the attention map that intersects with A (denoted as Attn_A, rendered in blue).

Mixup takes a pair of inputs $\mathbf{x}_A, \mathbf{x}_B$ and their corresponding labels $\mathbf{y}_A, \mathbf{y}_B$, then creates an artificial training example $\lambda\mathbf{x}_A + (1-\lambda)\mathbf{x}_B$ with $\lambda\mathbf{y}_A + (1-\lambda)\mathbf{y}_B$ as its ground truth. Here $\lambda \in [0,1]$ is the random mixing proportion sampled from a Beta distribution. This pre-assumes that linear interpolations of feature vectors should lead to linear interpolations of the associated targets.

However, we argue that the above pre-assumption does not always stay true since **not all pixels are created equal**. 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

*These authors contributed equally to this work.
Correspondence to Jie-Neng Chen (jienengchen01@gmail.com) and Shuyang Sun (kevinsun@robots.ox.ac.uk)

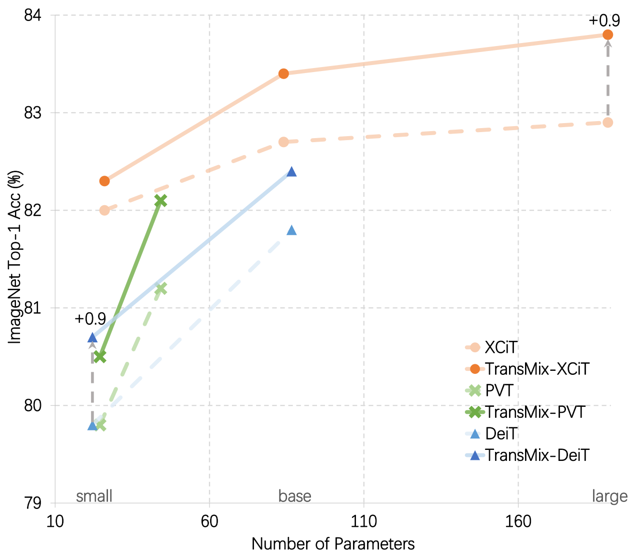

Figure 2. TransMix can steadily improve a wide range of state-of-the-art ViT-based models on ImageNet with no parameter and minimal computation overhead. See results for more model variants in Table 1.

gap between the input and the label space through the learning of label assignment. We find that the attention maps that are naturally generated in Vision Transformers can be well suited for this job. As shown in Figure 1, we simply set λ (weight of $\mathbf{y}_A$) as the sum of weights of attention map lying in A. In this way, the labels are re-weighted by the significance of each pixel instead of linearly interpolated with the same ratio as the mixed inputs. Since the attention map is naturally generated in ViT-based models, our method can be merged into the their training pipeline with no extra parameters and minimal computation overhead.

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 Fiugre 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 pretrained with TransMix on ImageNet, the superiority 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 rubust 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 re-assigns 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

$(\mathbf{x}_A, \mathbf{y}_A)$ and $(\mathbf{x}_B, \mathbf{y}_B)$ to augment a new training sample $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$. Formulaically,

$$\tilde{\mathbf{x}} = \mathbf{M} \odot \mathbf{x}_A + (\mathbf{1} - \mathbf{M}) \odot \mathbf{x}_B, \tag{1}$$

$$\tilde{\mathbf{y}} = \lambda \mathbf{y}_A + (1 - \lambda)\mathbf{y}_B, \tag{2}$$

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

During augmentation, a randomly sampled region in $\mathbf{x}_B$ is removed and filled in with the patch cropped from A of $\mathbf{x}_A$, where the patch's bounding box coordinates are uniformly sampled as (r_x, r_y, r_w, r_h). The mixed-target assignment factor λ 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 $\mathbf{x} \in \mathbb{R}^{N \times d}$, where N is the number of tokens, each of dimensionality d. The input $\mathbf{x}$ is linearly projected to queries, keys and values, using the weight matrices $\mathbf{w}_q \in \mathbb{R}^{d \times d_q}$, $\mathbf{w}_k \in \mathbb{R}^{d \times d_k}$ and $\mathbf{w}_v \in \mathbb{R}^{d \times d_v}$, such that $\mathbf{q}{=}\mathbf{x}\mathbf{w}_q$, $\mathbf{k}{=}\mathbf{x}\mathbf{w}_k$ and $\mathbf{v}{=}\mathbf{x}\mathbf{w}_v$, where $d_q = d_k$. Queries and keys are used to compute an attention map $\mathcal{A}(\mathbf{q}, \mathbf{k}) = \mathit{Softmax}(\mathbf{q}\mathbf{k}^\top / \sqrt{d_k}) \in \mathbb{R}^{N \times N}$, and the output of the self-attention operation is defined as the weighted sum of N token features in $\mathbf{v}$ with the weights corresponding to the attention map: $\text{Attention}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \mathcal{A}(\mathbf{q}, \mathbf{k})\mathbf{v}$. 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 and 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 $\mathbf{q}$ whose corresponding keys $\mathbf{k}$ are the all input tokens, and class attention $\mathbf{A}$ 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 $\mathbf{A}$ to mix labels.

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

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

$$\mathbf{q} = \mathbf{x}_{cls} \cdot \mathbf{w}_q, \tag{3}$$

$$\mathbf{k} = \mathbf{z} \cdot \mathbf{w}_k, \tag{4}$$

$$\mathbf{A}' = \mathit{Softmax}(\mathbf{q} \cdot \mathbf{k}^\top / \sqrt{d/g}), \tag{5}$$

$$\mathbf{A} = \{\mathbf{A}'_{0,i}, \mid i \in [1, p]\}, \tag{6}$$

where $\mathbf{q} \cdot \mathbf{k}^\top \in \mathbb{R}^{1 \times (1+p)}$ indicates the class token is a query whose corresponding keys are the all input tokens, and $\mathbf{A} \in [0,1]^p$ 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 $\mathbf{A} \in [0,1]^p$. In implementation, $\mathbf{A}$ 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 λ (the proportion of $\mathbf{y}_A$ in Eqn. (2)) with the guidance of the attention map $\mathbf{A}$:

$$\lambda = \mathbf{A} \cdot \downarrow(\mathbf{M}). \tag{7}$$

Here $\downarrow(\cdot)$ denotes the nearest-neighbor interpolation down-sampling that can transform the original $\mathbf{M}$ 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.

Algorithm 1 Pseudocode of TransMix in a PyTorch-like style.

```
# H, W: the height and width of the input image
# p: number of patches
# M: 0-initialized mask with shape (H,W)
# downsample: downsample from length (H*W) to (p)
# (bx1, bx2, by1, by2): bounding box coordinate

for (x, y) in loader: # load a minibatch with N pairs
    # CutMix image in a minibatch
    M[bx1:bx2, by1:by2] = 1
    x[:,:,M==1] = x.flip(0)[:,:,M==1]
    M = downsample(M.view(-1))

    # attention matrix A: (N, p)
    logits, A = model(x)

    # Mix labels with the attention map
    lam = matmul(A, M)
    y = (1-lam) * y + lam * y.flip(0)

    CrossEntropyLoss(logits, y).backward()
```

4.1. ImageNet Classification

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 hyperparameters compared to the DeiT [39] training recipe. We examined various baseline vision Transformer models including DeiT [39], PVT [48], CaiT [40], and XCiT [13], and the training schemes will be slightly adjusted 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 augment [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 $\pm 0.1\%$ compared with results reported in DeiT [39]. The attention map $\mathbf{A}$ 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 improvement 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.

Models	Params	FLOPs	Top-1 Acc (%)	+TransMix Top-1 Acc (%)
DeiT-T [39]	5.7M	1.6G	72.2	**72.6**
PVT-T [48]	13.2M	1.9G	75.1	**75.5**
XCiT-T [13]	12M	2.3G	79.4	**80.1**
CaiT-XXS[40]	17.3M	3.8G	79.1	**79.8**
DeiT-S [39]	22.1M	4.7G	79.8	**80.7**
PVT-S [48]	24.5M	3.8G	79.8	**80.5**
XCiT-S [13]	26M	4.8G	82.0	**82.3**
PVT-M [48]	44.2M	6.7G	81.2	**82.1**
PVT-L [48]	61.4M	9.8G	81.7	**82.4**
XCiT-M [13]	84M	16.2G	82.7	**83.4**
DeiT-B [39]	86.6M	17.6G	81.8	**82.4**
XCiT-L	189M	36.1G	82.9	**83.8**

Table 1. TransMix can steadily boost the a wide range of model variants *e.g*. DeiT, PVT, CaiT, and XCiT on ImageNet-1k classification. Note that all the baselines have been already carefully tuned with extensive augmentation and regularization techniques *e.g*. Mixup [54], CutMix[53], RandAug[10], DropPath[27] *etc*.

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 $z_{patches} \in \mathbb{R}^{p \times d}$ 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-

Backbone	Decoder	TransMix-pretrained	mAcc	mIoU	mIoU (MS)
ResNet101 [21]	Deeplabv3+ [8]		57.4	47.3	48.5
DeiT-S [39]	Linear		59.4	49.1	49.6
		✓	**60.2**	**49.7**	**50.3**
	Segmenter [37]		60.4	49.7	50.5
		✓	**61.4**	**50.6**	**51.2**

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.

Backbone	Params	Object detection			Instance segmentation		
		AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
ResNet50 [20]	44.2M	38.0	58.6	41.4	34.4	57.1	36.7
ResNet101 [20]	63.2M	40.4	61.1	44.2	36.4	57.7	38.8
PVT-S [48]	44.1M	40.4	62.9	43.8	37.8	60.1	40.3
TransMix-PVT-S	44.1M	**40.9**	**63.8**	**44.0**	**38.4**	**60.7**	**41.3**

Table 3. Overhead-free impact of TransMix on transferring to downstream **object detection and instance segmentation** using Mask R-CNN [20] with PVT [48] backbone on COCO val2017. AP^{bb} denotes bounding box AP for object detection and AP^{mk} denotes mask AP for instance segmentation.

coders: (1) Linear decoder (2) Segmenter decoder. 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 $z_{patches} \in \mathbb{R}^{p\times d}$ is used to produce patch-level logits $z_{lin} \in \mathbb{R}^{p\times K}$, 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 MMSegmentation [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 Object detection and instance segmentation experiments are conducted on COCO 2017. All models are trained on 118K images and evaluated 5K validation images. We study on PVT [48] as the detection backbone since its pyramid features make it favorable to object detection. The weights pre-trained on ImageNet is used to initialize the PVT backbone. We train and evaluate Mask R-CNN detector with the PVT backbone initialized with either vanilla (CutMix) or TransMix pre-trained weights for both object detection and instance segmentation. Following PVT [48], we adopt 1× training schedule (*i.e.*, 12 epochs) to train the detector on mmDetection [6] framework. Results for Mask R-CNN with ResNet backbone are reported in mmDetection [6] as references. As showed in Table 3, we find that without introducing extra parameter, the detector initialized with TransMix-pretrained backbone improves over CutMix-pretrained backbone by 0.5% box AP and 0.6% mask AP. Note that regularization-based pre-training for backbone has limited capability on improving downstream object detection. For instance, the recent Mixup variant SaliencyMix [41] only improved 0.16% box AP over CutMix-pretrained model on a smaller detection dataset.

4.3. Robustness Analysis

Recently the discussions regarding the robustness of vision Transformer are emerging [34, 32, 1]. To verify if TranMix 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 14x14 spatial grid; *i.e.* an image of size 224×224×3 is split into 196 patches, each of size 16×16×3. 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

	Nat. Adversarial Example			Out-of-Dist
Models	Top1-Acc	Calib-Error↓	AURRA	AUPR
DeiT-S	19.1	32.0	23.8	20.9
TransMix-DeiT-S	**21.1**	**31.2**	**28.8**	**21.9**

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

	Segmentation JI (%)	Localization mIoU (%)
DeiT-S	29.2	34.9
TransMix-DeiT-S	**29.9**	**44.4**

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.

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 $\mathbf{A}$ 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.

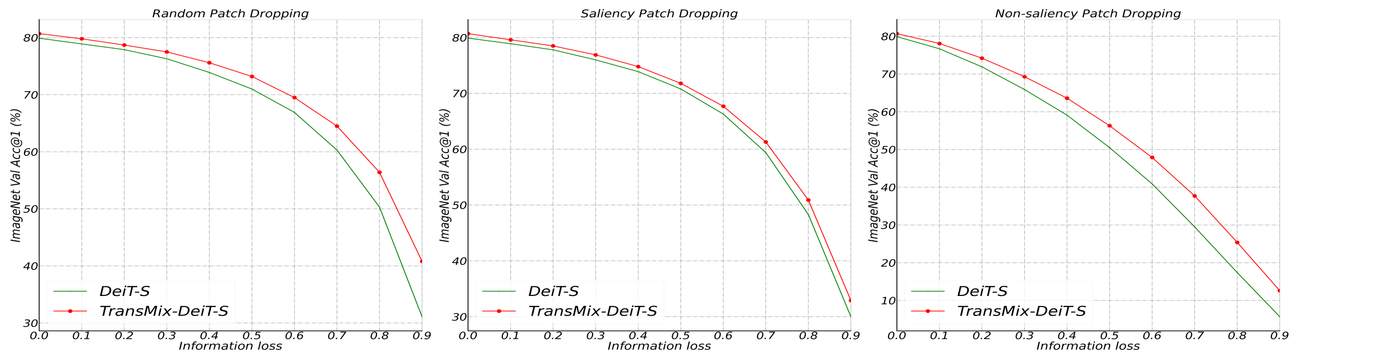

Figure 3. **Robustness against occlusion.** Model's robustness against occlusion with different information loss ratios is studied. 3 patch dropping settings: Random Patch Dropping (left), Salient Patch Dropping (middle), and Non-Salient Patch Dropping (right) are considered.

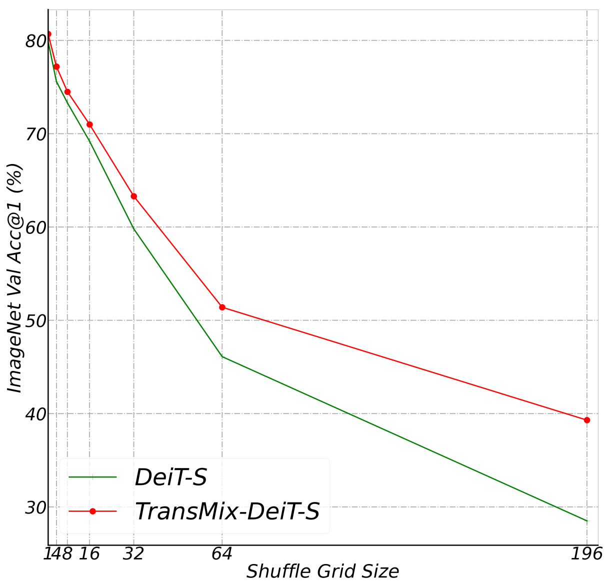

Figure 4. **Robustness against shuffle.** Model's robustness against shuffle with different grid shuffle sizes is studied. (Placeholder)

Attn Provider	Self	Dino	DeiT-pretrained	DeiT-distilled
top-1 Acc	80.7	80.6	80.4	80.4

Table 6. Using external (parameter-frozen) models to generate ttention map as the alternative to original attention map $\mathbf{A}$ used for TransMix.

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 $\mathbf{A}$.

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 $\mathbf{A}$ 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.

Models	Params	FLOPs	top-1 Acc (%)
Swin-T [30]	28.3M	4.5G	81.3
CA-Swin-T [30, 40]	28.3M	4.2G	81.6
TransMix-CA-Swin-T	28.3M	4.2G	**81.8**
Swin-S [30]	49.6M	8.8G	83.0
CA-Swin-S [30, 40]	49.6M	8.5G	82.8
TransMix-CA-Swin-S	49.6M	8.5G	**83.2**

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.

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

In this section, we provide the comprehensive comparison with many state-of-the-art mixup vairiants 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-

Method	Backbone	Params	Speed (im/sec)	top-1 Acc (%)
Baseline		22M	322	78.6
CutMix [53]		22M	322	79.8 (+1.2)
Attentive-CutMix [45]	DeiT-S	46M	239	77.5 (-1.1)
SaliencyMix [41]		22M	314	79.2 (+0.6)
Puzzle-Mix [28]		22M	139	79.8 (+1.2)
TransMix		22M	322	**80.7** (+2.1)

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.

vided in the supplementary material. All mentioned models are built upon DeiT training recipe towards a fair comparison. Baseline in Table 8 is chosen to be the default DeiT-S framework excluding CutMix in training. Measured on image per second (im/sec), training speed (*i.e.* training throughput) is performed in average of five runs for images at resolution 224×224 under 128 batch size with a Tesla-V100 graphic card, and 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).

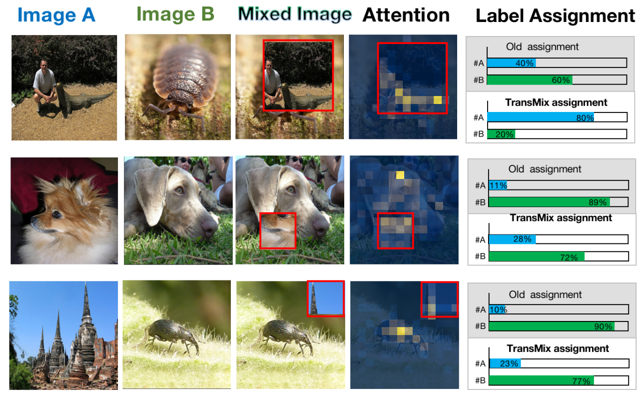

Figure 5. The visualization including image A, image B, mixed image, attention map obtained from XCiT-L when input mixed image, and corresponding label assignments. The label assignments include both the old area-ratio assignment and new TransMix assignment.

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

[1] Yutong Bai, Jieru Mei, Alan Yuille, and Cihang Xie. Are transformers more robust than cnns? In *NeurIPS*, 2021. 5

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2021. 6

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6

[4] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *ICCV*, 2021. 4

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 5

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 5

[9] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020. 5

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 4

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009. 4

[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 *ICLR*, 2020. 1, 2, 3

[13] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. In *NeurIPS*, 2021. 1, 2, 4

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. `http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html`, 2012. 6

[15] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. In *NeurIPS*, 2021. 1, 2

[16] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[17] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. Keepaugment: A simple information-preserving data augmentation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1055–1064, 2021. 2

[18] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, and Alan Yuille. Transfg: A transformer architecture for fine-grained recognition. In *AAAI*, 2022. 2

[19] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4918–4927, 2019. 4

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 5

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[22] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *ICLR*, 2020. 2

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, 2021. 6

[24] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 4

[25] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8129–8138, 2020. 4

[26] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661. Springer, 2016. 4

[27] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661. Springer, 2016. 4

[28] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *ICML*, pages 5275–5285, 2020. 1, 2, 8

[29] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *NeurIPS*, 2019. 6

[30] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 7

[31] Fan Lu, Guang Chen, Yinlong Liu, Lijun Zhang, Sanqing Qu, Shu Liu, and Rongqi Gu. Hregnet: A hierarchical network for large-scale outdoor lidar point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16014–16023, 2021. 2

[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

[33] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5

[34] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *NeurIPS*, 2021. 5, 6

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 6

[36] Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. *AAAI*, 2022. 2

[37] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5

[38] Shuyang Sun, Liang Chen, Gregory Slabaugh, and Philip Torr. Learning to sample the most useful training patches from images. *arXiv preprint arXiv:2011.12097*, 2020. 2

[39] 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*, pages 10347–10357, 2021. 1, 2, 4, 5

[40] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021. 1, 2, 4, 7

[41] AFM Uddin, Mst Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae, et al. Saliencymix: A saliency guided data augmentation strategy for better regularization. In *ICLR*, 2021. 1, 2, 5, 8

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 3

[44] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *ICML*, pages 6438–6447, 2019. 2

[45] Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. *arXiv preprint arXiv:2003.13048*, 2020. 1, 2, 8

[46] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5463–5474, 2021. 5

[47] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9695–9704, 2021. 2

[48] 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

[49] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 4

[50] Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan Yuille, and Wei Shen. Glance-and-gaze vision transformer. In *NeurIPS*, 2021. 2

[51] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021. 4

[52] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip HS Torr, Wayne Zhang, and Dahua Lin. Vision transformer with progressive sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 387–396, 2021. 8

[53] 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

[54] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1, 2, 4

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 8