# Supplementary Materials: "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.

## A. Additional Experimental Details

### A.1. More Details to Train Classification Models

We examined various baseline vision Transformer models including DeiT [13], PVT [17], CaiT [14], XCiT [6], and Swin [10]. As we try to carry out a nearly unified training scheme for different models, we make minimal changes to hyperparameters compared to the DeiT [13] training recipe, unless specified otherwise. In doing so, the training schemes will be slightly adjusted to the official implementations of individual model variants.

We primarily follow the settings of the data augmentation and regularization adopted in [13], including RandAug [4], Stochastic Depth [8], Mixup [19] and CutMix [18]. We don't adopt repeated augment [7]. For all models, the initial learning rate, the total training batch size, and weight decays is 0.001, 1000, 0.03 respectively. We set warmed up for 20 epochs expect DeiT-B keeping 5 epochs to reach the initial learning rate . All Transformers are trained for 300 epochs expect that El-Nouby et al. [6] and Touvron et al. [14] report 400 epochs for XCiT and CaiT respectively. The accuracy of our baseline implementation fluctuates only by $\pm 0.1\%$ compared with results reported in DeiT [13]. Note that we do not use any external dataset for pre-training and we do not use knowledge distillation.

**DeiT**

### A.2. Implementation Details of Compared Mixup Variants

The comparison with state-of-the-art Mixup variants is conducted in Section 4.6. We explain the implementation details here. The official implementations of Mixup variants are mainly based on the backbone of ResNet-50, and we apply their methods into training DeiT-S.

**Baseline** Baseline in Table 8 is chosen to be the default DeiT-S framework excluding CutMix in training.

**Attentive-CutMix** Attentive-CutMix is implemented based on the unofficial pytorch repository [1]. Attentive-CutMix contains an affiliated model (i.e. ResNet-50) for saliency map extraction and a backbone model for image classification.

[1]https://github.com/xden2331/attentive_cutmix

| $d$ | 6 | 8 | 10 | 12 | rollout |
|---|---|---|---|---|---|
| top-1 Acc | 80.3 | 80.3 | 80.4 | 80.7 | 80.4 |

Table 1. **Ablation study** on attention generation. Attention matrix used for TransMix is output from the $d$-th block of DeiT-S. Following [1, 2], rollout applies matrix multiplication across all 12 blocks' attention matrices.

**SaliencyMix** Saliency-Mix is implemented based on the official pytorch codebase [2]. SaliencyMix uses third-party library opencv to extract the saliency map with

```
cv2.saliency.StaticSaliencyFineGrained_create()
```

**Puzzle-Mix** Puzzle-Mix is implemented following the official pytorch codebase [3]. Puzzle-Mix forwards and backwards the model twice to detect object saliency by computing the gradients of the neural network following [12].

## B. Additional Results

**Ablation Study** The class attention $\mathbf{A}$ can obtained from any Tranformer Block in ViTs. Due to the global receptive field, the class attention would not have big difference across blocks [11, 5]. We first study the effect of attention matrix generated in different depth $d$ for DeiT-S. Then we follow [1, 2] to compute the attention rollout, which aggregate the attention matrices from all blocks by matrix multiplications. According to the results, we found that the default setting with $d = 12$ performs the best. Notably, the total number of Transformer block with class token is varying in different vision Transformers (*e.g.* 24 for XCiT, 2 for CaiT, 12 for DeiT). Particularly, PVT designs hierarchical Transformer blocks with 4 different resolution scales, and therefore an extra downsample step is a must if using early scale attention matrices. Hence, using the attention from the last Transformer block as default can not only avoid finding a optimal $d$ exhaustingly but also be compatible for all ViT variants.

[2]https://github.com/afm-shahab-uddin/SaliencyMix
[3]https://github.com/snu-mllab/PuzzleMix

| Method | Backbone | Params | top-1 Acc (%) |
|---|---|---|---|
| Baseline | ResNet-50 | 25M | 76.3 |
| CutMix [18] | | 25M | 78.6 |
| SaliencyMix [15] | | 25M | 78.7 |
| Puzzle-Mix [9] | | 25M | 78.8 |
| Baseline | DeiT-S | 22M | 78.6 |
| CutMix [18] | | 22M | 79.8 |
| Attentive-CutMix [16] | | 46M | 77.5 |
| SaliencyMix [15] | | 22M | 79.2 |
| Puzzle-Mix [9] | | 22M | 79.8 |
| TransMix | | 22M | 80.7 |

Table 2. Comparison with state-of-the-art Mixup variants with the backbone of either ViT or CNN on ImageNet-1k. All listed models are trained for 300 epochs towards fair comparison. ResNet-50 results are borrowed from the paper [15].

**Results on DeiT with knowledge distillation** DeiT's accuracy can be further boosted with knowledge distillation [13]. Here we conduct experiments on DeiT distillation, and TransMix can improve DeiT-S-Distill and DeiT-B-Distill without cost consistently. TransMix lifts the accuracy of DeiT-S-Distill from 81.2% to 81.6%, and the accuracy of DeiT-B-Distill from 83.4% to 83.7%.

**Mixup variants on CNN and ViT** We also attach the official results of some Mixup variants with the backbone of CNN. Results on the ResNet-50 backbone are borrowed from [9]. All models are trained for 300 epochs towards fair comparison. As backbone, DeiT-S has similar number of parameters to ResNet-50. Table 2 shows that SaliencyMix and Puzzle-Mix only improve over CutMix by at most 0.2% on ResNet-50 and show no advancement over CutMix on DeiT-S.

## C. More Visualizations

We provide more visualizations as shown in Figure 1.

**Effects of Different Augmentations** Following [13], we conduct ablation study on different types of strong data augmentation including Random-Augment [4], Auto-Augment [3], Mixup [19], Cutmix [18] and our TransMix. The ablation study is evaluated on the model of DeiT-S on ImageNet-1k.

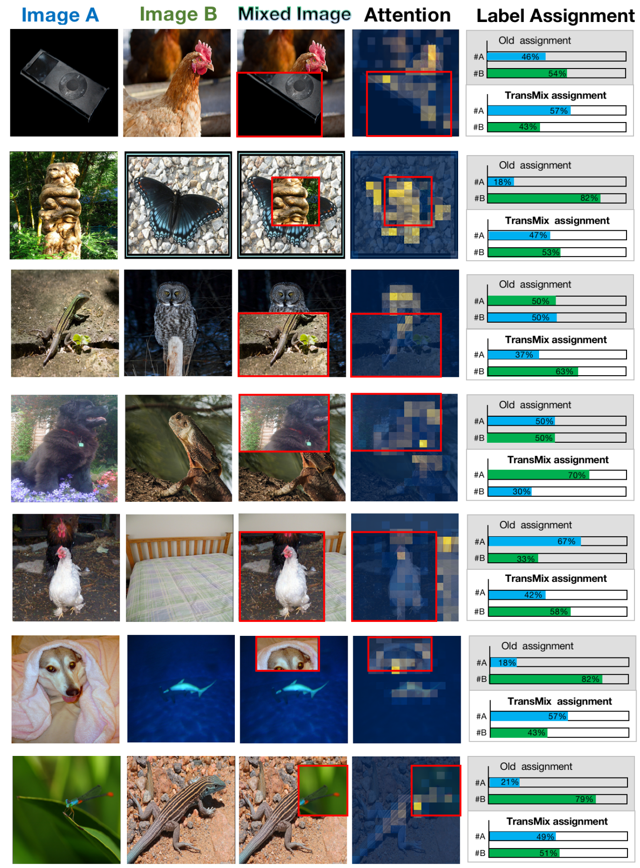


Figure 1. The visualization including image A, image B, mixed image, attention map obtained from XCiT-L when input mixed image, and corresponding label assignments. The label assignments include both the old area-ratio assignment and new TransMix assignment.

| BaseAug | RandAug | Mixup | CutMix | TransMix | top-1 Acc (%) |
|---|---|---|---|---|---|
| ✗ | ✗ | ✗ | ✗ | ✗ | 57 |
| ✓ | ✗ | ✗ | ✗ | ✗ | 73.3 |
| ✓ | ✓ | ✗ | ✗ | ✗ | 76.5 |
| ✓ | ✓ | ✓ | ✗ | ✗ | 78.6 |
| ✓ | ✓ | ✓ | ✓ | ✗ | 79.9 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 80.7 |

Table 3. Ablation study on augmentation strategy for DeiT-S on ImageNet-1k. The symbols ✓ and ✗ indicate that we use and do not use the corresponding augmentations, respectively.

## References


[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, pages 4190–4197, 2020. 1

[2] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, pages 782–791, June 2021. 1

[3] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. 2019. 2

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmenta-

tion with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 1, 2

[5] 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

[6] 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

[7] 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. 1

[8] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661. Springer, 2016. 1

[9] 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. 2

[10] 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

[11] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *NeurIPS*, 2021. 1

[12] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014. 1

[13] 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

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

[15] 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. 2

[16] 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. 2

[17] 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

[18] 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

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