Token-Label Alignment for Vision Transformers
Supplementary Material

Han Xiao[1,2,*] Wenzhao Zheng[1,2,*] Zheng Zhu[3] Jie Zhou[1,2] Jiwen Lu[1,2,†]
[1]Beijing National Research Center for Information Science and Technology, China
[2]Department of Automation, Tsinghua University, China [3]PhiGent Robotics
{h-xiao20,zhengwz18}@mails.tsinghua.edu.cn; zhengzhu@ieee.org;
{jzhou,lujiwen}@tsinghua.edu.cn

A. Comparisons of Different Training Recipes

We compare different training recipes for the DeiT-S model in Table 1. The results of TransMix [2] reported in the original paper adopts an advanced training recipe with a model exponential moving average, resulting in slower training speed. Differently, we basically follow the conventional DeiT-S [17] training recipe and improve its performance by 0.8%. We report the result of TransMix with the same training recipe (80.1%) in Table 2 of the main text.

B. Details of Experimental Analysis

Details about Datasets We evaluate our method on ImageNet [13] for image classification, ADE20K [19] for semantic segmentation, and COCO 2017 [10] for object detection and instance segmentation. ImageNet [13] contains about 1.2 million training and 50K validation images from 1K categories. ADE20K [19] contains 20K training images and 2K validation images from 150 semantic categories. COCO 2017 [10] dataset consists of 118K training images and 5K validation images from 80 different categories. We further conduct experiments to evaluate the robustness and the generalization ability of the TL- Align pretrained models. For robustness, we consider ImageNet-A [9], ImageNet-C [8], ImageNet-R [7], and under AutoAttack [3]. ImageNet-A [9] consists of naturally adversarial examples from real-world challenging scenarios. ImageNet-C [8] is used to evaluate the model robustness to diverse image corruptions. ImageNet-R [7] contains various artistic renditions of 200 ImageNet classes. which contains new test sets of ImageNet following the same labeling protocol. AutoAttack [3] is a novel adversarial attacks benchmark to test the adversarial robustness on ImageNet validation set. To evaluate the generalization ability, we adopt the ImageNet-V2 dataset [12] which contains new test sets of ImageNet following the same labeling protocol.

Obtaining the "Ground-truth" Mixing Ratio. To better illustrate the token fluctuation phenomenon, we compute a "ground-truth" mixing ratio based on token similarity as shown in Figure 1. Formally, given two input images $\mathbf{X}_1$, $\mathbf{X}_2$ and their mixed sample $\tilde{\mathbf{X}}$ generated by CutMix, we feed all of them into the vision transformer to obtain the corresponding tokens $\mathbf{Z}_1^l$, $\mathbf{Z}_2^l$ and $\tilde{\mathbf{Z}}^l$ after the transformer block l. For each mixed token $\tilde{\mathbf{z}}_i^l$ in $\tilde{\mathbf{Z}}^l$, we compute its maximum cosine similarity with all tokens in $\mathbf{Z}_1^l$ and $\mathbf{Z}_2^l$, respectively, as:

$$\mathrm{s}_1^l(\tilde{\mathbf{z}}_i^l) = \max_j \frac{(\tilde{\mathbf{z}}_i^l)^T \mathbf{z}_{1j}^l}{||\tilde{\mathbf{z}}_i^l|| \cdot ||\mathbf{z}_{1j}^l||}, \quad \mathrm{s}_2^l(\tilde{\mathbf{z}}_i^l) = \max_j \frac{(\tilde{\mathbf{z}}_i^l)^T \mathbf{z}_{2k}^l}{||\tilde{\mathbf{z}}_i^l|| \cdot ||\mathbf{z}_{2k}^l||} \tag{1}$$

The contribution of input $\mathbf{X}_1$ to the token $\tilde{\mathbf{z}}_i^l$ is then obtained using the softmax function: $\lambda = \mathrm{softmax}(\mathrm{s}_1^l(\tilde{\mathbf{z}}_i^l), \mathrm{s}_2^l(\tilde{\mathbf{z}}_i^l))$. We visualize this similarity-based mixing ratio of the class token in DeiT-S in Figure 4 of the main text. As shown, the token mixing ratio changes after processing by each transformer block, which demonstrates the token fluctuation problem. Moreover, TL-Align assigns a dynamic mixing ratio to tokens at different layers, which is more consistent with the "ground truth" compared with other methods. This provides an empirical analysis to explain the improvement achieved by our TL-Align.

Implementation of Different Data Mixing Strategies. We provide implementation details of different data mixing strategies that we adopt to evaluate the effectiveness of TL-Align. Inspired by MAE [6] and BEiT [1], we implement a random mixing strategy and block-wise mixing strategy. The visualization of the mixed images produced by CutMix, random mixing, and block-wise mixing strategies is shown in Figure 2. Specifically, employing the block-wise strategy leads to a top-1 accuracy of 80.0%, which is the high-

[*] Equal contribution.
[†] Corresponding author.

Table 1. **Comparisons of different training recipes for the DeiT-S model on ImageNet-1K.**

Method	Training Epochs	Warmup Epochs	LR	Weight Decay	Model EMA	EMA Decay	MixUp	CutMix	MixUp Switch Prob	Random Erasing	Top-1 Acc. (%)
DeiT-S[1] [17]	300	5	0.0005	0.05	×	-	0.0	0.0	-	✓	76.4
DeiT-S[2] [17]	300	5	0.0005	0.05	×	-	0.8	1.0	0.5	✓	79.8
DeiT-S[3] [17]	310	20	0.001	0.03	✓	0.99996	0.8	1.0	0.8	×	80.3
+TransMix [2]	310	20	0.001	0.03	✓	0.99996	0.8	1.0	0.8	×	80.7 (+0.4)
DeiT-S[4] [17]	300	5	0.0005	0.05	×	-	0.0	1.0	-	✓	79.8
+TransMix [2]	300	5	0.0005	0.05	×	-	0.0	1.0	-	✓	80.1 (+0.3)
+TL-Align	300	5	0.0005	0.05	×	-	0.0	1.0	-	✓	**80.6 (+0.8)**

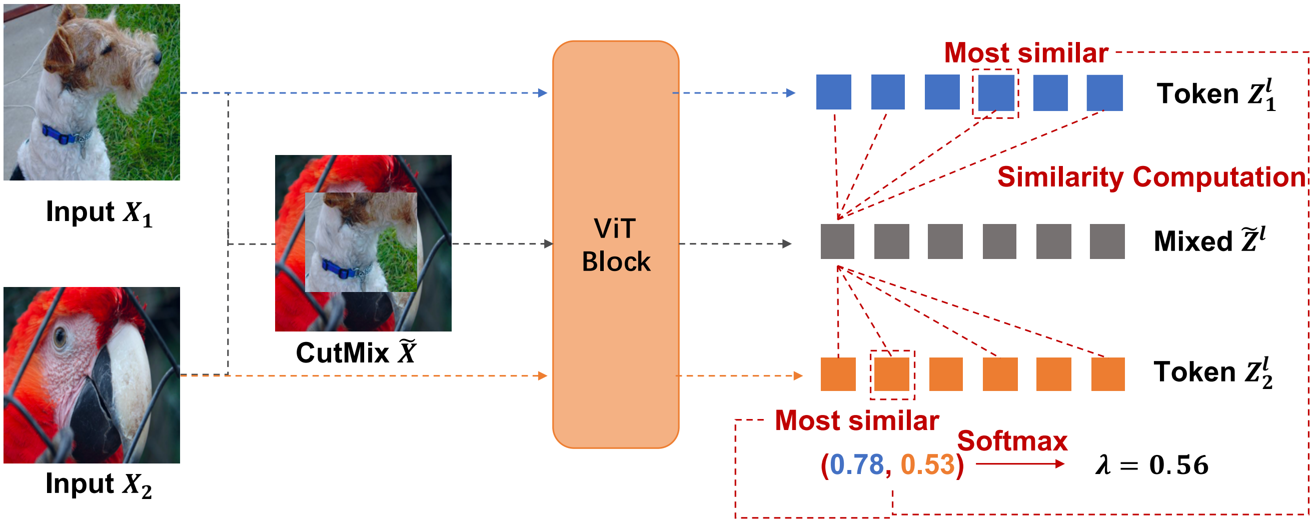

Figure 1. Illustration of how we get a "ground-truth" mixing ratio based on token similarity.

est among the data mixing strategies. Our TL-Align further boosts the accuracy by $+0.3\%$, verifying its generalizability on various data mixing strategies.

C. More Visualization Results

We provide more visualization results of the obtained labels by the proposed TL-Align in Figure 3. We visualize the input images, the mixed image, the original label embedding, and the label embedding after our TL-Align. Specifically, we visualize the aligned label embedding after the final transformer block for both DeiT-S and Swin-S. The size of the original label embedding is equivalent to the number of input tokens, i.e., 14×14 for DeiT-S and 56×56 for Swin-Transformer since they employ different patch sizes for patch embedding. The value of the label embedding represents the probability of which class the corresponding token belongs to, which is shown by color. Red stands for the class of the first input image while blue stands for the class of the second input image. We observe that the aligned labels can deviate from the original labels, resulting in different mixing ratios during training. Therefore, using the original mixing ratio as the training target produces false training signals and might lead to inferior performance.

D. Generalizing TL-Align Beyond ViTs

ViTs can achieve better accuracy/computation trade-off than conventional CNNs, where one of the working mechanisms is the alternation between spatial mixing (e.g., SA) and channel mixing (e.g., MLP) [15]. Based on this, some works have explored different spatial mixing strategies in addition to self-attention, including spatial MLP [15, 16, 14, 18] and depth-wise convolution [4, 11, 5]. For an image $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, they first perform patch-wise image tokenization to obtain a tokenized image representation $\mathbf{Z} \in \mathbb{R}^{N\times d}$, where N is the number of tokens and d is the number of channels. To generalize TL-Align to other architectures beyond ViTs, we first formulate modern deep vision networks into various compositions of five operations:

- Spatial mixing: $\mathbf{Z} \leftarrow \mathbf{W}^s(\mathbf{Z}) \cdot \mathbf{Z}$, where $\mathbf{W}^s(\mathbf{Z}) \in \mathbb{R}^{N\times N}$.
- Channel mixing: $\mathbf{Z} \leftarrow \mathbf{Z} \cdot \mathbf{W}^c(\mathbf{Z})$, where $\mathbf{W}^c(\mathbf{Z}) \in \mathbb{R}^{d\times d}$.
- Point-wise transformation: $\mathbf{Z} \leftarrow f(\mathbf{Z})$, where f is a point-wise operation such as bias adding and normalization.

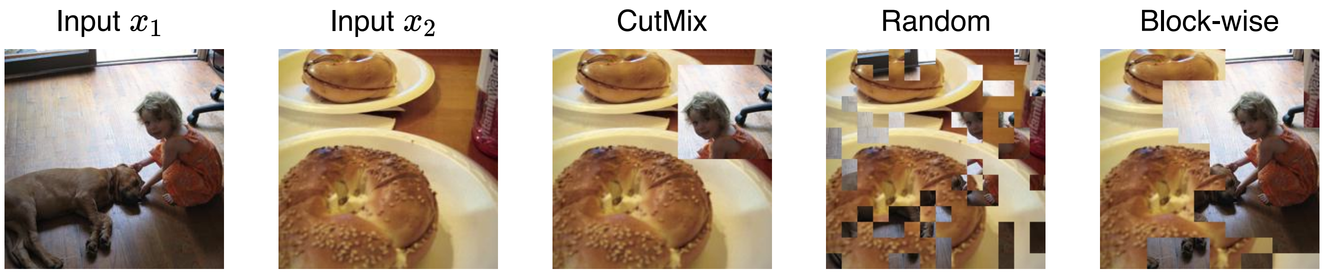

Figure 2. Visualization of mixed images produced by different data mixing strategies.

Table 2. **Updating of the label embeddings for different operations on the tokens.**

Operation	Token Processing	Label Alignment	Example
Spatial mixing	$\mathbf{Z} \leftarrow \mathbf{W}^s(\mathbf{Z}) \cdot \mathbf{Z}$	$\mathbf{Y} \leftarrow \text{Norm}(\mathbf{W}^s(\mathbf{Z})) \cdot \mathbf{Y}$	Spatial attention
Channel mixing	$\mathbf{Z} \leftarrow \mathbf{Z} \cdot \mathbf{W}^c(\mathbf{Z})$	$\mathbf{Y} \leftarrow \mathbf{Y}$	Channel MLP
Point-wise transformation	$\mathbf{Z} \leftarrow f(\mathbf{Z})$	$\mathbf{Y} \leftarrow \mathbf{Y}$	Layer normalization
Residual connection	$\mathbf{Z} \leftarrow \mathbf{Z} + g(\mathbf{Z})$	$\mathbf{Y} \leftarrow \text{Norm}(\mathbf{Y} + g(\mathbf{Y}))$	Residual connection
Spatial aggregation	$\mathbf{Z} \leftarrow \text{Aggre}(\{\mathbf{Z}_i\})$	$\mathbf{Y} \leftarrow \text{Norm}(\sum_i \mathbf{Y}_i)$	Patch merging

- Residual connection: $\mathbf{Z} \leftarrow \mathbf{Z} + g(\mathbf{Z})$, where g can be one or a composition of the aforementioned operations.
- Spatial aggregation: $\mathbf{Z} \leftarrow \text{Aggre}(\{\mathbf{Z}_i\})$, where Aggre typically concatenates multiple tokens across the feature dimension.

For example, MLP-Mixer [15] adopts $\mathbf{W}^s(\mathbf{Z}) = W^s$, where $W^s \in \mathbb{R}^{N \times N}$ is a learnable parameter matrix. ConvNeXt [11] adopts $\mathbf{W}^s(\mathbf{Z}) = T(\mathbf{K})$, where $\mathbf{K} \in \mathbb{R}^{7 \times 7}$ is a convolutional kernel and T transforms the kernel into an equivalent matrix for direct multiplication.

The proposed TL-Align can be generalized to different architectures by applying the corresponding operations on the label embeddings. We initialize the label embedding following Eq. 5 in the main text. We detail the label embedding updating for different operations in Table 2. The Norm$(\cdot)$ operation denotes that we normalize each row vector so that the sum of all elements equals to 1.

For spatial mixing, we accordingly mix the token embeddings using the same weights as the token processing. For example, for a processed token $\hat{\mathbf{z}} = 0.5 \cdot \mathbf{z}_1 + 0.5 \cdot \mathbf{z}_2$, we similarly compute the aligned label as $\hat{\mathbf{y}} = 0.5 \cdot \mathbf{y}_1 + 0.5 \cdot \mathbf{y}_2$, assuming the label information is linearly addable. As channel mixing and point-wise transformation only reorganize information within each token, they do not alter the label embedding. For residual connection, we similarly add a residual connection to the label embedding before normalization. Spatial aggregation is similar to spatial mixing and also aggregates information among multiple tokens. Therefore, we also need to align the labels by adding their label embeddings before normalization. We leave the experiments for generalized TL-Align for future works.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1

[2] Jie-Neng Chen, Shuyang Sun, Ju He, Philip Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. *arXiv preprint arXiv:2111.09833*, 2021. 1, 2

[3] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 1

[4] Xiaohan Ding, Xiangyu Zhang, Yizhuang Zhou, Jungong Han, Guiguang Ding, and Jian Sun. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *CVPR*, 2022. 2

[5] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *arXiv preprint arXiv:2202.09741*, 2022. 2

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[7] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1

[8] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1

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

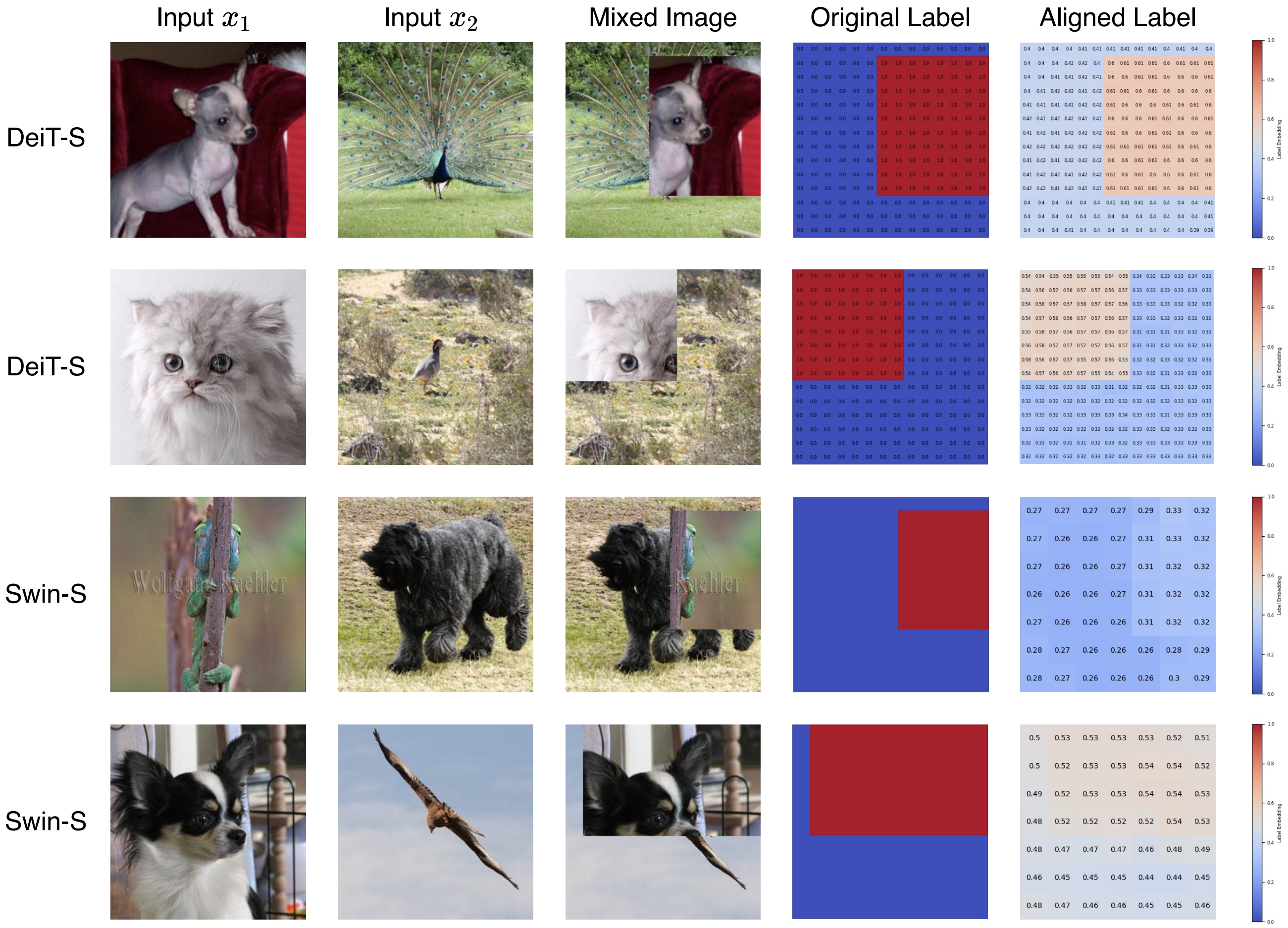

Figure 3. More visualization results on DeiT-S and Swin-S. We visualize the input images, the mixed image, the original label embedding, and the label embedding after token-label alignment.

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15262–15271, 2021. 1

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1

[11] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022. 2, 3

[12] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1

[13] 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. *IJCV*, 115(3):211–252, 2015. 1

[14] Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Yanxi Li, Chao Xu, and Yunhe Wang. An image patch is a wave: Quantum inspired vision mlp. In *CVPR*, 2022. 2

[15] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *NeurIPS*, 34, 2021. 2, 3

[16] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021. 2

[17] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357, 2021. 1, 2

[18] Guoqiang Wei, Zhizheng Zhang, Cuiling Lan, Yan Lu, and Zhibo Chen. Activemlp: An mlp-like architecture with active token mixer. *arXiv preprint arXiv:2203.06108*, 2022. 2

[19] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127:302–321, 2019. 1