A1. Statistics of Attention Score

Attention score can reflect the similarity of each token to the others. For mixed images, tokens from the same unmixed image are more similar than those from different unmixed images. To further corroborate the visualization in Figure 5c of the main paper with quantitative data, we calculate average attention scores among image tokens from different regions of the mixed images.

How to calculate? We leverage SMMix to generate new mixed images based on ImageNet-1k [9]. A mixed image contains two regions respectively from the source and target images. The mixed image is fed into a ViT to obtain a token sequence, $\mathbf{T} \in \mathbb{R}^{N \times d}$. For a simpler representation, we simply assume that $\mathbf{I}_s = \{\mathbf{I}_s^1, \mathbf{I}_s^2, ..., \mathbf{I}_s^{N_s}\}$ and $\mathbf{I}_t = \{\mathbf{I}_t^1, \mathbf{I}_t^2, ..., \mathbf{I}_t^{N_t}\}$, where $\mathbf{I}_s$ and $\mathbf{I}_t$ are the indexes of the tokens from the source and target regions, respectively; N_s and N_t respectively indicate the token number of the source and target regions, and $N_s + N_t = N$. Following Eq. (6) in the main paper, we obtain the self-attention matrix, $\mathbf{A} \in \mathbb{R}^{N \times N}$, which contains attention scores among each token; $\mathbf{A}^{i,j}$ denotes the attention score when taking the i-th token as a query and the j-th token as a key. There are two types of tokens, either from the source or target region. Thus, self-attention forms four (query, key) pairs for mixed images according to the token region. Table A1 shows how to calculate average attention scores for the four (query, key) pairs.

Results. We can find two interesting phenomena in Table A2:

First, SMMix assists tokens focus more on tokens from the same regions. For example, when both the query and key tokens are from the same regions, the SMMix pre-trained model has attention scores of 0.0142 and 0.0070, which are higher than the CutMix pre-trained model's 0.0122 and 0.0046.

Second, SMMix alleviates incorrect attention scores caused by sharp rectangle boundaries. Taking tokens from target regions as queries, we find that the CutMix pre-trained model focuses more on tokens from source regions (0.0098) than tokens from target regions (0.0046). The incorrect attention scores are caused by sharp rectangle boundaries, which enhance the first/second-order feature statistics and cause self-attention operation to generate basic attention scores for the cropped rectangles regardless of contents. However, taking tokens from target regions as a query, SMMix pre-trained models successfully focus more on tokens from target regions (0.0070), rather than tokens from source regions (0.0021).

These two phenomena show that ViTs pre-trained with SMMix can generate more appropriate attention scores and help the model locate the accurate regions.

A2. Additional Results

Comparisons with TokenLabel. Tabel A3 compares our SMMix with TokenLabel [26]. We observe that SMMix outperforms TokenLabel in DeiT-T (+0.7%) and DeiT-S (+0.1%). Also, SMMix has less training time and without dependence on any pre-trained models, while TokenLabel requires a NAFNet-F6 model [1] that has 438M parameters.

Variants of max-min attention region mixing. For the max-min attention region mixing, we select the maximum-scored region from a source image and paste it to the minimum-scored region in a target image. Such an operation can maximize the information of mixed images and make the proposed fine-grained label assignment feasible. To demonstrate the effectiveness of such a mixing pattern, we consider five possible variants:

- (i) Random $\rightarrow$ Corr: randomly select a region from the source image and paste it to the same location in the target image;
- (ii) Random $\rightarrow$ Max Attn: randomly select a region from the source image and pastes it to the maximum-scored region in the target image;
- (iii) Random $\rightarrow$ Min Attn: which randomly select a region from the source image and paste it to the minimum-scored region in the target image;
- (iv) Max Attn $\rightarrow$ Corr: which select the maximum-scored region from the source image and paste it to the same location in the target image;
- (v) Max Attn $\rightarrow$ Max Attn: which select the maximum-scored region from the source image and paste it to the maximum-scored region in the target image.

Finally, we denote our max-min attention region mixing as Max Attn $\rightarrow$ Min Attn. Table A4 compares the performance. Obviously, our $Max\ Attn \rightarrow Min\ Attn$ achieves the best performance compared to its variants, because it maximizes the information of mixed images. On the other hand, $Random \rightarrow Max\ Attn$ performs the worst, since it occludes the most targets. Note that these findings are inconsistent with SalinencyMix [46], which believes that $Attn \rightarrow Corr$ pattern performs best since the pattern provides a trade-off between regularization and image information. We attribute the difference to two possible causes: (1) Our image attention score locates objects more accurately than the salience detector in SaliencyMix [46]; (2) The regularizations strategies in the ViTs training recipe allow more information to be retained in mixing methods.

Image Attention Score. Table A5 shows the performance for image attention scores from different depths. We

Query \ Key	Source	Target
Source	$\frac{1}{N_s^2}\sum \mathbf{A}^{i,j}, s.t.\ i \in \mathbf{I}_s, j \in \mathbf{I}_s$	$\frac{1}{N_s N_t}\sum \mathbf{A}^{i,j}, s.t.\ i \in \mathbf{I}_s, j \in \mathbf{I}_t$
Target	$\frac{1}{N_s N_t}\sum \mathbf{A}^{i,j}, s.t.\ i \in \mathbf{I}_t, j \in \mathbf{I}_s$	$\frac{1}{N_t^2}\sum \mathbf{A}^{i,j}, s.t.\ i \in \mathbf{I}_t, j \in \mathbf{I}_t$

Table A1: Calculation detail of average attention scores between image tokens from different regions.

Table A2: Attention scores among image tokens from source/target regions. Intuitively, tokens should pay more attention to the tokens from the same regions. Score1/**Score2** refers to corresponding attention scores of the models trained with CutMix and SMMix, respectively.

Query \ Key	Source	Target
Source	0.0122/**0.0142**↑	0.0037/**0.0031**↓
Target	0.0098/**0.0021**↓	0.0046/**0.0070**↑

Table A3: Comaprison of our SMMix with TokenLabel on ImageNet-1k. "Pra-trained" indicates whether to adopt a pre-trained model for the network training. "Time" refers to the training time increase over CutMix.

Model	Method	Pre-trained	Time	Top-1 Acc.(%)
DeiT-T [44]	Baseline	✗	1.00×	72.2
	TokenLabel	✓	1.59×	72.9
	SMMix (ours)	✗	1.10×	**73.6**
DeiT-S [44]	Baseline	✗	1.00×	79.8
	TokenLabel	✓	1.59×	81.0
	SMMix (ours)	✗	1.10×	**81.1**

Table A4: Ablation of different image mixing schemes on DeiT-S. All the models are trained for 100 epochs.

Mixing Scheme	Top-1 Acc.(%)
Random → Corr	74.2
Random → Max Attn	73.8
Random→ Min Attn	74.5
Max Attn → Corr	74.4
Max Attn → Max Attn	74.3
Max Attn → Min Attn(Ours)	**74.7**

observe 0.4% performance drop when taking a random image attention score, demonstrating the guidance ability of the image attention score in the image mixing process. SMMix achieves the best performance when $d = 6, 9, 12$. We set $d = 12$ as the default since the feature consistency constraint requires a complete forward propagation. However, this shows that when the feature consistency constraint is disabled, SMMix can further reduce training costs by using the shallower-layer image attention score.

Table A5: Ablation of image attention score generation. SM-Mix uses the image attention score output by the d-th block of DeiT-S. "None" means to randomly generate the image attention score. "Rollout" means to average all-block image attention scores.

d	None	3	6	9	12	Rollout
Top-1 Acc.(%)	80.7	81.0	81.1	81.1	81.1	81.1

A3. Details of Training Time Testing

In Figure 1 of the main paper, we report the training time of DeiT-S [44] on different CutMix variants. All models are trained on ImageNet-1k with a 4×A100 GPU machine for 300 epochs, and AMP [38] is activated during the training process. In particular, we follow the original DeiT training recipe except for TransMix [4]. Following the open source code of TransMix [4], we reproduce it by modifying the batch size from 1024 to 256.

A4. More Visualization

Figure A1 and Figure A2 provide more visual examples in ImageNet-1k. The visualization shows that models trained with our SMMix can locate objects more accurately in both unmixed and mixed images.

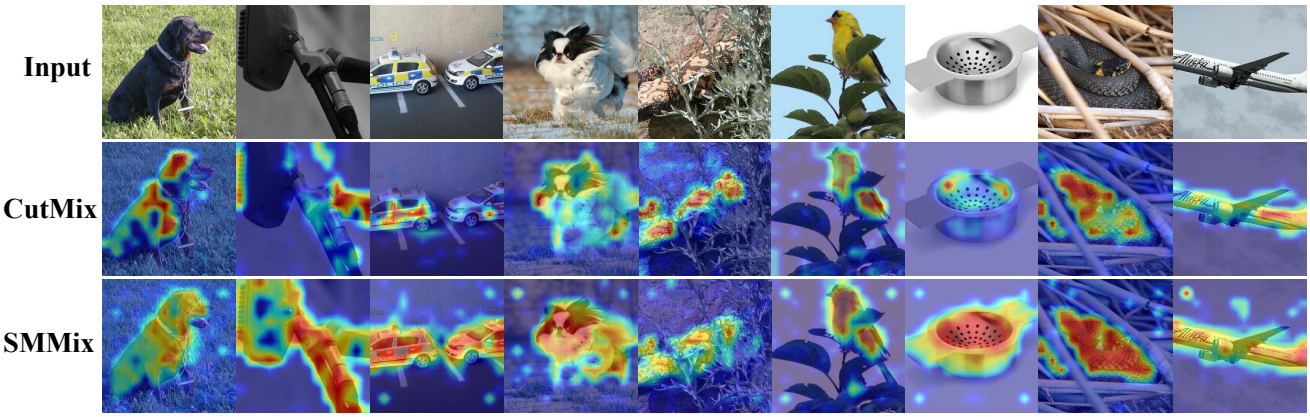

Figure A1: The class activation map [41] of the models trained with CutMix and SMMix and tested on unmixed images.

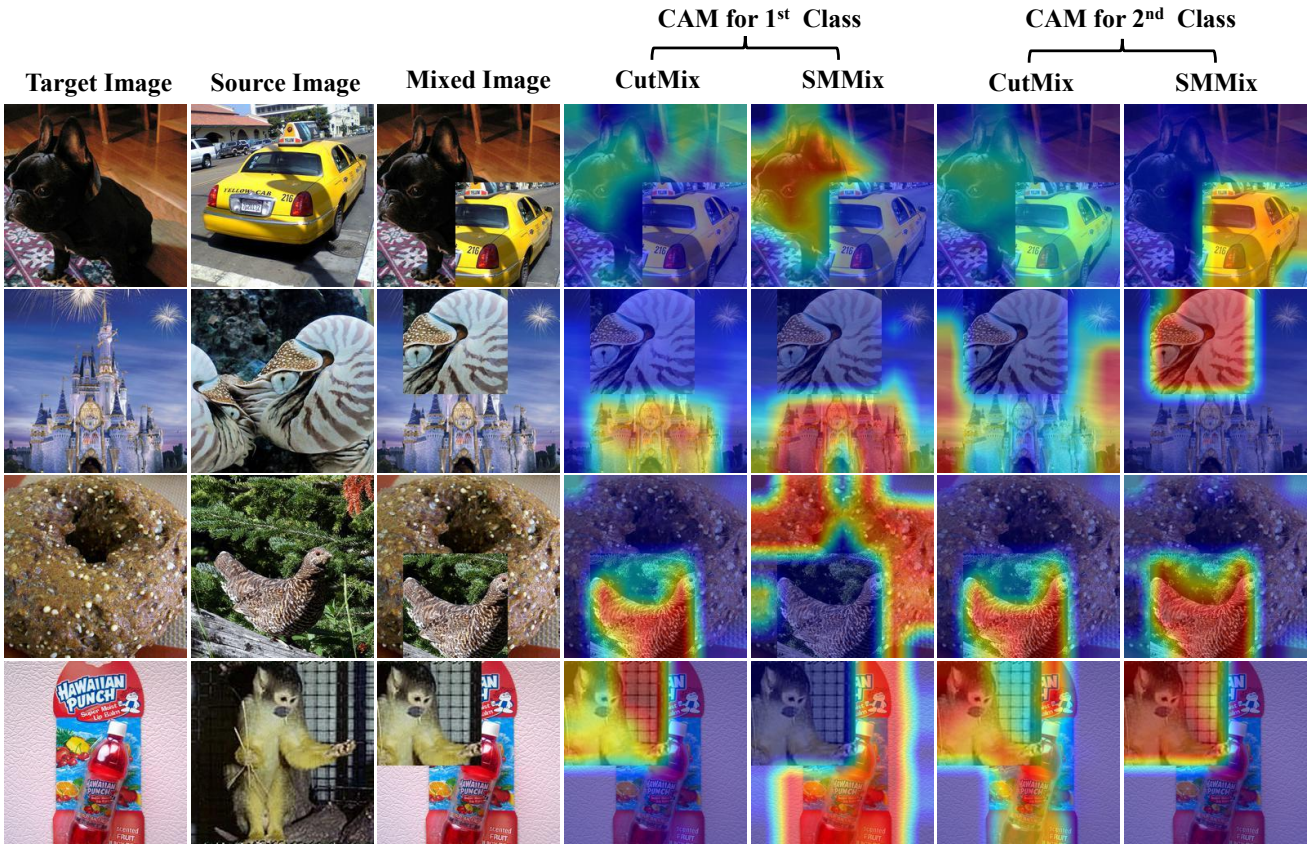

Figure A2: The class activation map [41] of the models trained with CutMix and SMMix and tested on mixed images.