A.Appendix

A.1 Details about aligning ViT's dense predictions to CNN's dense predictions.

In this part, we provide the details about aligning dense predictions. The shape for ViT's dense predictions is $(1, N, C)$; and the shape of CNN's dense predictions is $(1, H, W, C)$. Classical CNNs have a downsampling ratio of 32 while ViT takes 16×16 sub-images as patches. Thus, the number of ViT tokens is 4 times the number of CNN's dense responses ($N = 2H \times 2W$). Given a 224×224 image, the last feature map of a classicial CNN is in shape 7×7 and the number of patch tokens is 196 (14×14). Hence, it's needed to align the spatial dimensions of the two responses. In CSKD, we simply use a 2×2 average pooling (the stride is set as 2) to downsample ViT's responses. We first reshape ViT's responses to $(1, 14, 14, C)$, then feed them to the pooling layer to generate the aligned responses with the shape of $(1, 7, 7, C)$. Thus, The shapes of CNN's dense predictions and ViT's dense predictions are *exactly the same*. We further calculate the KL-Divergence or cross-entropy loss on the aligned responses.

A.2 Ensemble of patch tokens.

Since CSKD also trains the patch tokens, it is convenient to utilize them for classification ensemble. In DeiT, the output of distillation token is combined with the output of class token for the final prediction. In CSKD, the outputs of all patch tokens are further combined. Concretely, we calculate the prediction results of all tokens. All patch token outputs are averaged for the final "patch token output". Then, we average the "patch token output", class token output, and distillation token output as the final prediction. Experiments show that it can further improve the performance by 0.1%. It indicates that the patch tokens are useful for more accurate classification, but the performance is still excellent without extra computation cost (*i.e.*, without ensemble).

	ensemble	top-1
CSKD-Ti		76.1
	✓	76.3
CSKD-S		82.2
	✓	82.3
CSKD-B		83.7
	✓	83.8

Table 10. Caption

A.3 Implementations of transfer learning.

We follow the training setting of DeiT's repository[7]. We fine-tune all models with 384×384 resolution on the downstream datasets (without high-resolution fine-tuning on ImageNet-1k).

A.4 More visualizations.

We provide more visualizations of the attention heatmaps.

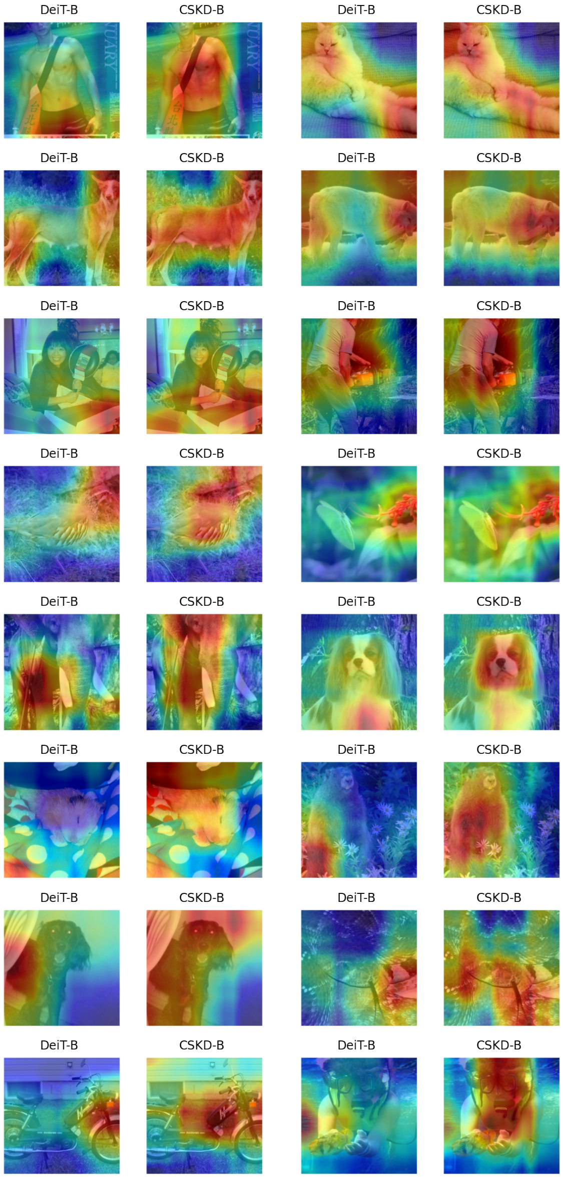

Figure 7. Attention Heatmaps from DeiT-B and our CSKD-B. CSKD-B focuses more attention on the salient object.

[7] https://github.com/facebookresearch/deit/issues/105