

# Appendix

November 21, 2021

## 1 EXPERIMENTS

### 1.1 EARLY STOP EPOCH

Table 1 shows the performance with different early stop epochs. It can be seen that stopping distillation at 100 epoch can achieve the best results.

Model	SL KD	SSL KD	Early Stop Epoch	Head Sel.	Acc@1
Baseline	×	×	×	-	72.2
SSTA_KD_early50	✓	✓	50	imp.	73.4
SSTA_KD_early100 ( <b>Ours</b> )	✓	✓	100	imp.	<b>74.0</b>
SSTA_KD_early150	✓	✓	150	imp.	73.4
SSTA_KD_early200	✓	✓	200	imp.	73.7
SSTA_KD_early250	✓	✓	250	imp.	72.6
SSTA_KD_early300	✓	✓	×	imp.	72.2

Table 1: Ablation study of early stop epoch on DeiT-Ti. imp. stands for selecting the most important head for distillation.

### 1.2 LAYERS FOR DISTILLATION

As the representation similarity between SL teacher and SSL teacher (SSTA) is lower in the deeper layers (see Figure 2 in the paper), which means the diversity between SL teacher and SSL teacher (SSTA) is higher, we search the layers from back to front. It can be observed that with the increase of the number of distillation layers, the accuracy of the student rises first, when the number of layers is 3 (*i.e.*  $L = 10, 11, 12$ ), it reaches the maximum value, and then if the number of layers increases again, the accuracy will decrease instead. Therefore, our distillation is carried out on the 10th, 11th and 12th layers.

Layers	ACC@1
{12}	73.1
{11, 12}	73.4
{10, 11, 12}	<b>74.0</b>
{9, 10, 11, 12}	73.6

Table 2: The effect of distilling with different layers on ImageNet-1K.

### 1.3 SHAPE BIAS

The results of shape bias are presented in Figure 1, we can see that although the shape bias of SL teacher is higher than that of SSL, the shape bias of student distilled by two different SL teachers is actually lower than that of student distilled by one SL teacher and together with another SSL teacher (SSTA). **The results manifest that our proposed SSTA lets students have a higher shape bias and behave more like human.**

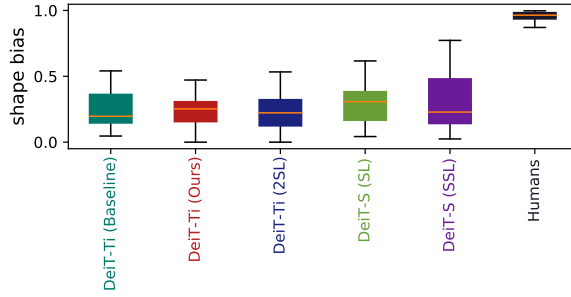


Figure 1: Shape bias of ViTs. DeiT-S (SL) and DeiT-S (SSL) are two teachers and DeiT-Ti (2SL) is distilled by two different SL teachers. The horizontal line in each rectangular entity is the median.

## 2 VISUALIZATIONS OF THE LAST LAYER

In Figure 2, we present some visualizations of the attention map from the last layer. As our proposed SSTA transfers knowledge on the first head of student, we can see that the first head pay more attention to objects, compared to the baseline.

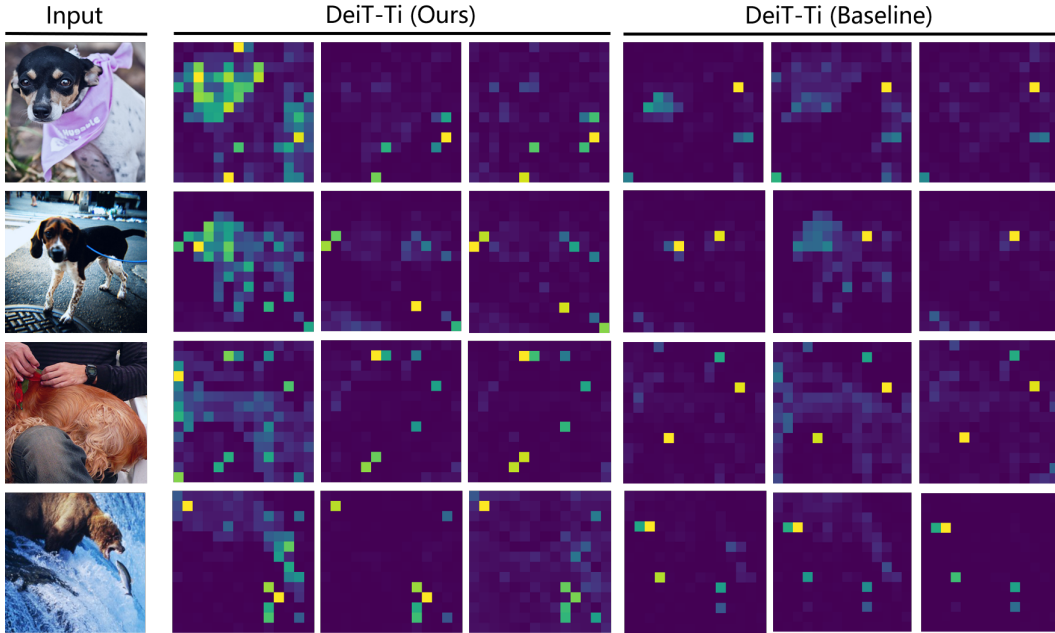


Figure 2: Visualizations of self-attention map from the last layer. There are three heads on DeiT-Ti.

## 3 VISUALIZATIONS OF THE MOST IMPORTANT HEAD.

For the proposed head selection strategy, we provide the visualizations of the most important heads of the last 3 layers of SL teacher and SSL teacher (SSTA) in Figure 3. We can see that the background is more important for SL teacher, while the information of object is more critical for the SSL teacher (SSTA). Actually, background and object both are important for

human vision, thus our head-level knowledge distillation method precisely adopt the most critical information via head selection strategy.

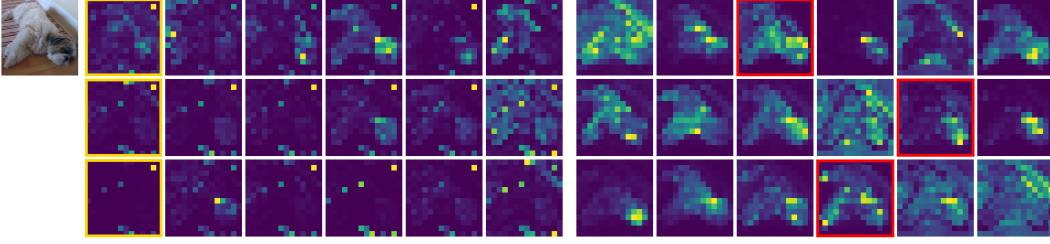


Figure 3: The first column is the input image, the next 6 columns are the self-attention visualization of the 6 heads of SL teacher, and the last 6 columns belong to SSL teacher (SSTA). The first row to the third row correspond to the 10th, 11th, and 12th layers respectively. The heads with boxes are the most important heads of corresponding teachers over three layers.

#### 4 VISUALIZATIONS OF DISTILLED HEADS ON STUDENT

Figure 4 shows the visualizations of the student (DeiT-Ti) after mimicking SL teacher and SSTA. Since the baseline model can not pay attention to the object ('tin opener') precisely and disturbed by redundant information, the object is identified as a pencil sharpener. On the contrary, our SSL Teacher (SSTA) perfectly focus on the object and the heads that selected by the proposed head selection strategy provide the critical information to the student. After distilling, DeiT-Ti can also precisely focus on the object and classify it correctly.

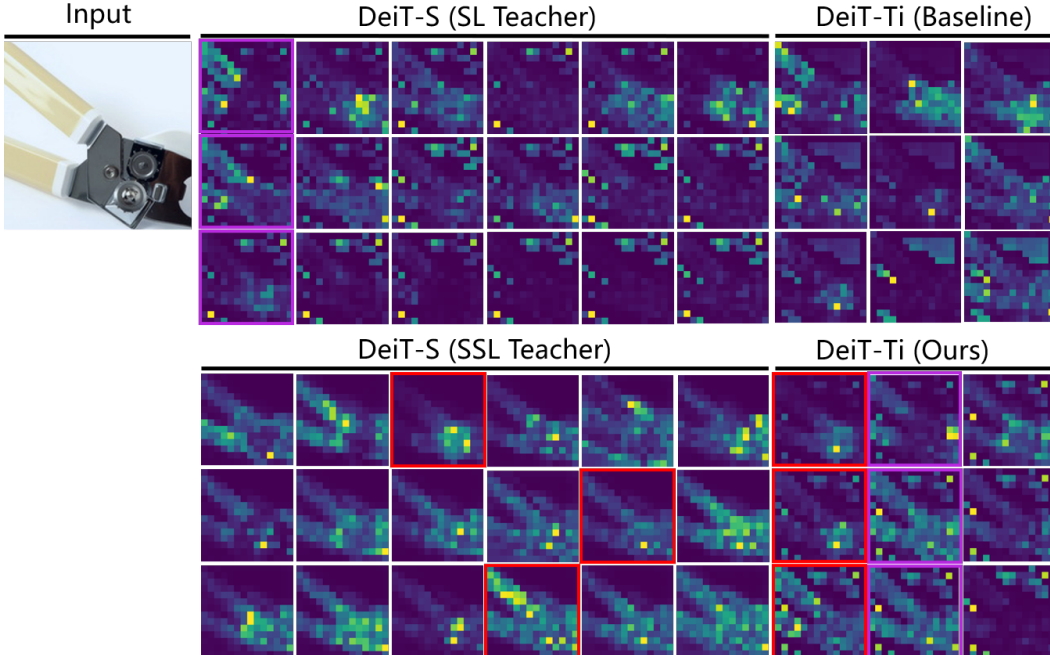


Figure 4: Visualizations of self-attention from the last 3 layers of two teachers, our student and baseline. The red boxes and purple boxes on teachers denote the most important head of SSL teacher (SSTA) and SL teacher. Meanwhile, the red boxes and purple boxes on student denote the heads distilled by the SSL teacher (SSTA) and SL teacher correspondingly.