A IMPLEMENTATION DETAILS

Through our layer-to-phase training, we determine that three pruning phase is best for computation-accuracy tradeoff. For DeiT-T/S/B, we insert the token selector before the 3rd, 6th, and 9th layer. For LV-ViT-S, we insert the token selector before the 4th, 8th, and 12th layer. For LV-ViT-M, we insert the token selector before the 5th, 10th, and 15th layer. For PiT-T/XS/S, we insert the token selector before the 1st, 5th, and 10th layer Our searched pruning rate for each phase is 0.383, 0.631, 0.863, respectively. The throughput is measured on a single NVIDIA A100-SXM4-40GB GPU with batch size of 256.

B MORE ANALYSIS

B.1 TOKEN PRUNING VISUALIZATION

We further visualize the process of HFSP to describe the performance in the inference phase and make a comparison between the framework with the token packaging technique and without it the token packaging technique. As shown in Figure 4, the first row shows the results collected from the framework without token packaging and the second row is from the framework with token packaging. At phase 1, there is a 7% difference in the left image groups and 11% in the right ones; At phase 2, 11% difference in the left image groups and 15% in the right ones; At phase 3, 14% difference in the left image groups and 18% in the right ones. We can infer token packaging can help to lock the object instead of the background. And this phenomenon is more obvious in complex and multi-object images.

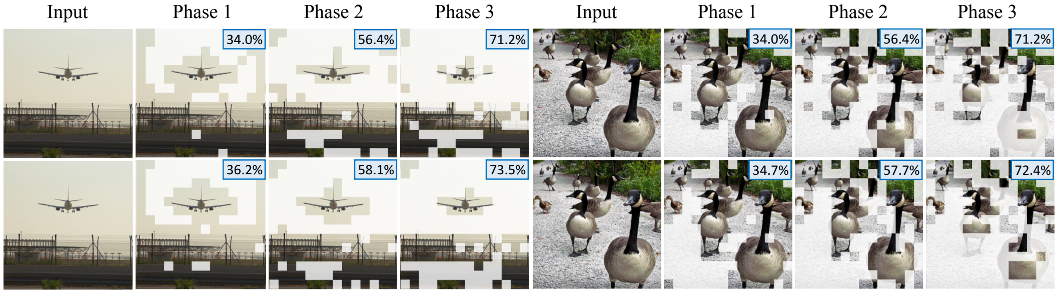

Figure 4: Visualization of each pruning phase. The first row shows the results collected from the framework without token packaging and the second row shows the results from the framework with token packaging. Tool refers to (Rao et al., 2021).

B.2 NUMBER OF PACKAGE TOKENS

For all models, we insert three soft pruning modules for hierarchical pruning. In each module, a new package token is generated. Here we conduct two ways to pass on the package token for subsequent layers: 1) Merge the package token generated from the current module with the existing one from the last module by element-wise addition. Therefore, only 1 additional token is added to the input sequence in total. 2) Concatenate the new package token to the existing ones. Making it 3 additional tokens added in total. Table 7 shows a comparison of the two methods.

Table 7: Package token preservation number comparison on DeiT-S

Method	**Params (M)**	**FLOPs (G)**	**Top-1 Acc. (%)**	**Top-5 Acc. (%)**
Baseline	22.10	4.60	79.80	95.00
1 × Package Token	22.20	2.63	79.26	94.55
+ Attention-based Branch	22.20	2.64	79.30	94.60
3 × Package Token	22.20	2.64	79.28	94.65
+ Attention-based Branch	22.20	2.64	79.34	94.67

B.3 DIFFERENT BATCH SIZE COMPARISON

We run our HFSP on DeiT-S with different batch size for ablation. Results in Table 8 show that accuracy has a slight boost when batch size increases.

Table 8: Different Batch Size Comparison on DeiT-S

Batch Size	**Params (M)**	**FLOPs (G)**	**Top-1 Acc. (%)**	**Top-5 Acc. (%)**
96	22.20	2.65	79.31	94.64
128	22.20	2.65	79.32	94.64
256	22.20	2.64	79.34	94.67

B.4 SELF-ATTENTION HEAD HEATMAP

Figure 5 shows the heatmaps of informative region detected by each self-attention head in DeiT-S. Each attention head focuses on encoding different image features and visual receptive fields. Therefore, token importance is different for each head. This demonstrates the need for obtaining token score individually.

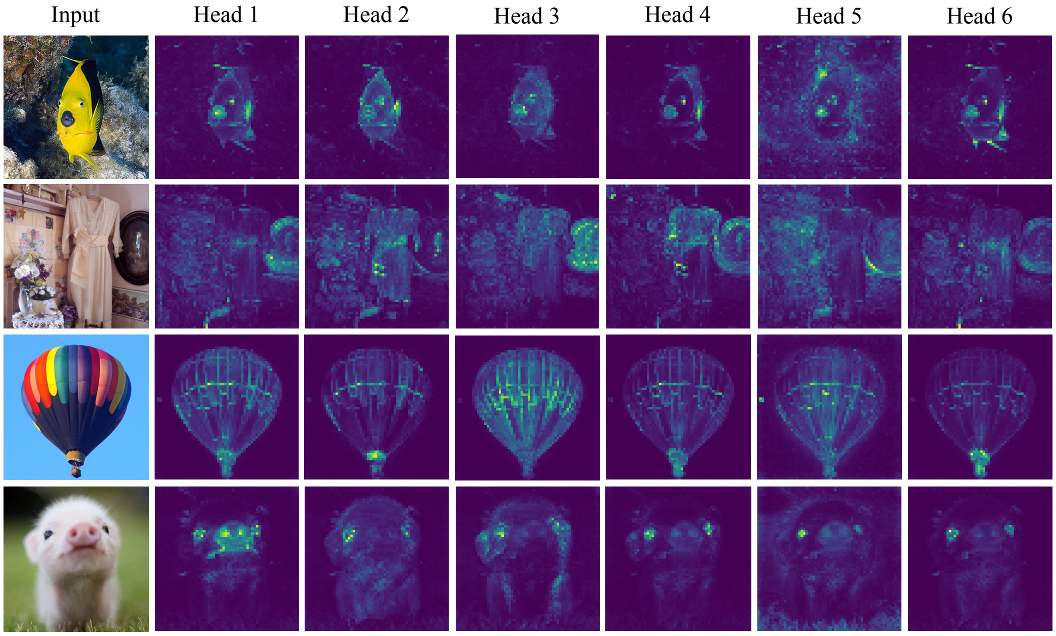

Figure 5: Heatmaps showing the informative region detected by each head in DeiT-S.

B.5 THE CORRELATION COEFFICIENT OF THE TOKEN FEATURES IN EACH LAYER

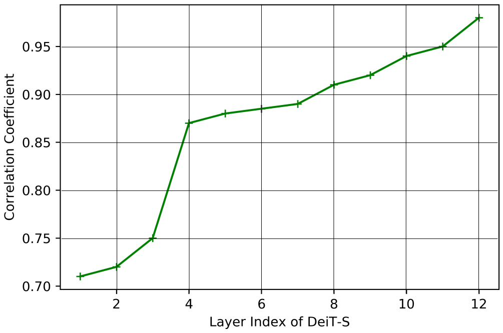

Figure 6: The correlation coefficient of the token features in each layer of DeiT-S.

As shown in Figure 6, the lower similarities with larger variation in the shallow layers shows that the token representations are insufficiently encoded, and it is difficulty to recognize the redundancies in shallow features.