AdaViT: Adaptive Vision Transformers for Efficient Image Recognition
Appendix

Lingchen Meng[1,2,3*] Hengduo Li[4*] Bor-Chun Chen[5] Shiyi Lan[4]
Zuxuan Wu[1,2†] Yu-Gang Jiang[1,2] Ser-Nam Lim[5]
[1]Shanghai Key Lab of Intelligent Info. Processing, School of Computer Science, Fudan University
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3]Biren Technology [4]University of Maryland [5]Meta AI

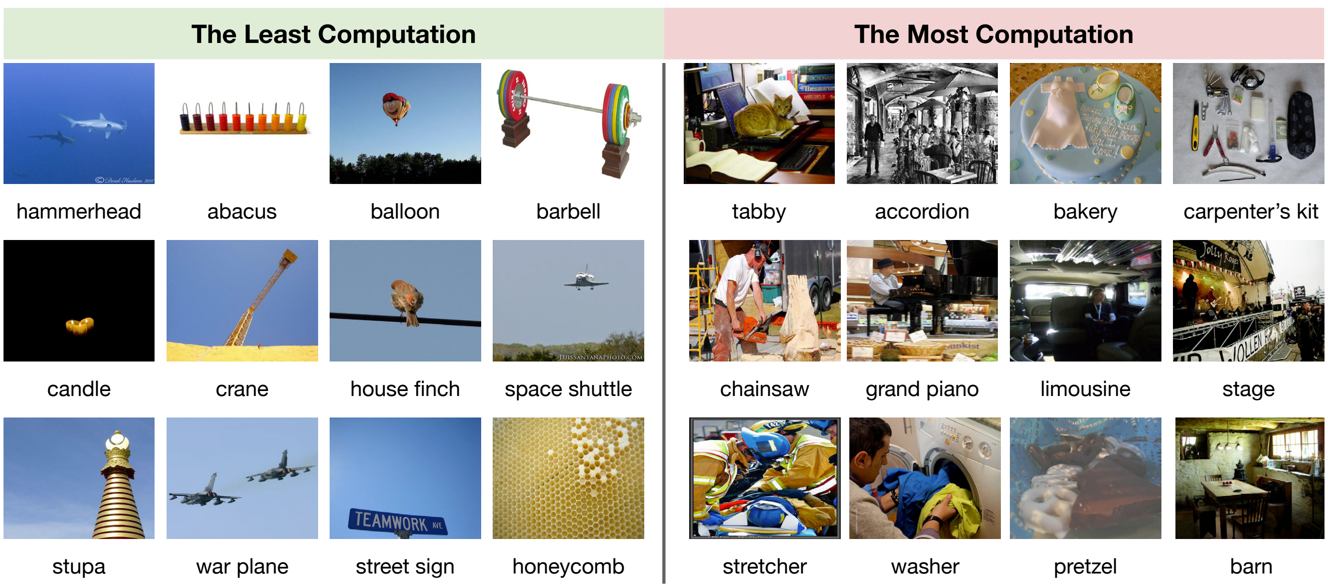

Figure 1. **Qualitative results.** Images allocated with the least (**Left**) and the most (**Right**) computational resources by AdaViT are shown.

A. Qualitative Results

We further provide more qualitative results in addition to those in the main text. Images that are allocated the least/most computational resources by our method are shown in Figure 1, demonstrating that our method learns to use less computation on easy object-centric images and more computation on hard complex images with cluttered background. Figure 3 shows more visualization of the learned usage policies for patch selection, demonstrating the pattern that our method allocates less and less computation gradually throughout the backbone network, which indicates that more redundancy in computation resides in the later stages of the vision transformer backbone.

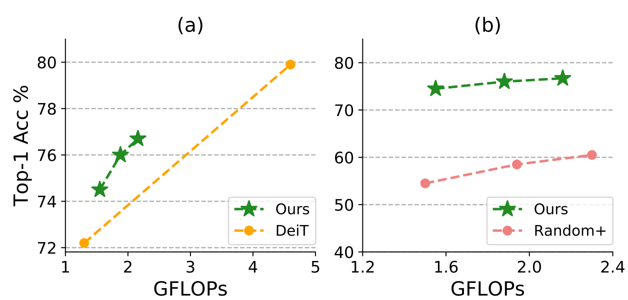

Figure 2. **Compatibility to DeiT [1].** We use DeiT-small as the backbone of AdaViT and show: **(a)** Efficiency/Accuracy tradeoffs of standard DeiT variants and our AdaViT. **(b)** Comparison between AdaViT and its *Random*+ baseline with similar computational cost.

B. Compatibility to Other Backbones

Our method is by design model-agnostic and thus can be applied to different vision transformer backbones. To verify this, we use DeiT-small [1] as the backbone of AdaViT and show the results in Figure 2. AdaViT achieves better efficiency/accuracy tradeoff when compared with standard

Figure 3. **Visualization of selected patches at different blocks** with T2T-ViT [2] (**Above**) or DeiT [1] (**Below**) as the vision transformer backbone respectively. Green color denotes the patch is kept.

variants of DeiT, and consistently outperforms its *Random+* baseline by large margins, as demonstrated in Figure 2(a) and 2(b) respectively.

We further show the visualization of patch selection usage policies with DeiT-small as the backbone as well in Figure 3. A similar trend of keeping more computation at earlier layers and gradually allocating less computation throughout the network is also observed.

References

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

[2] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng

Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, 2021. 2