Appendix A - More Examples

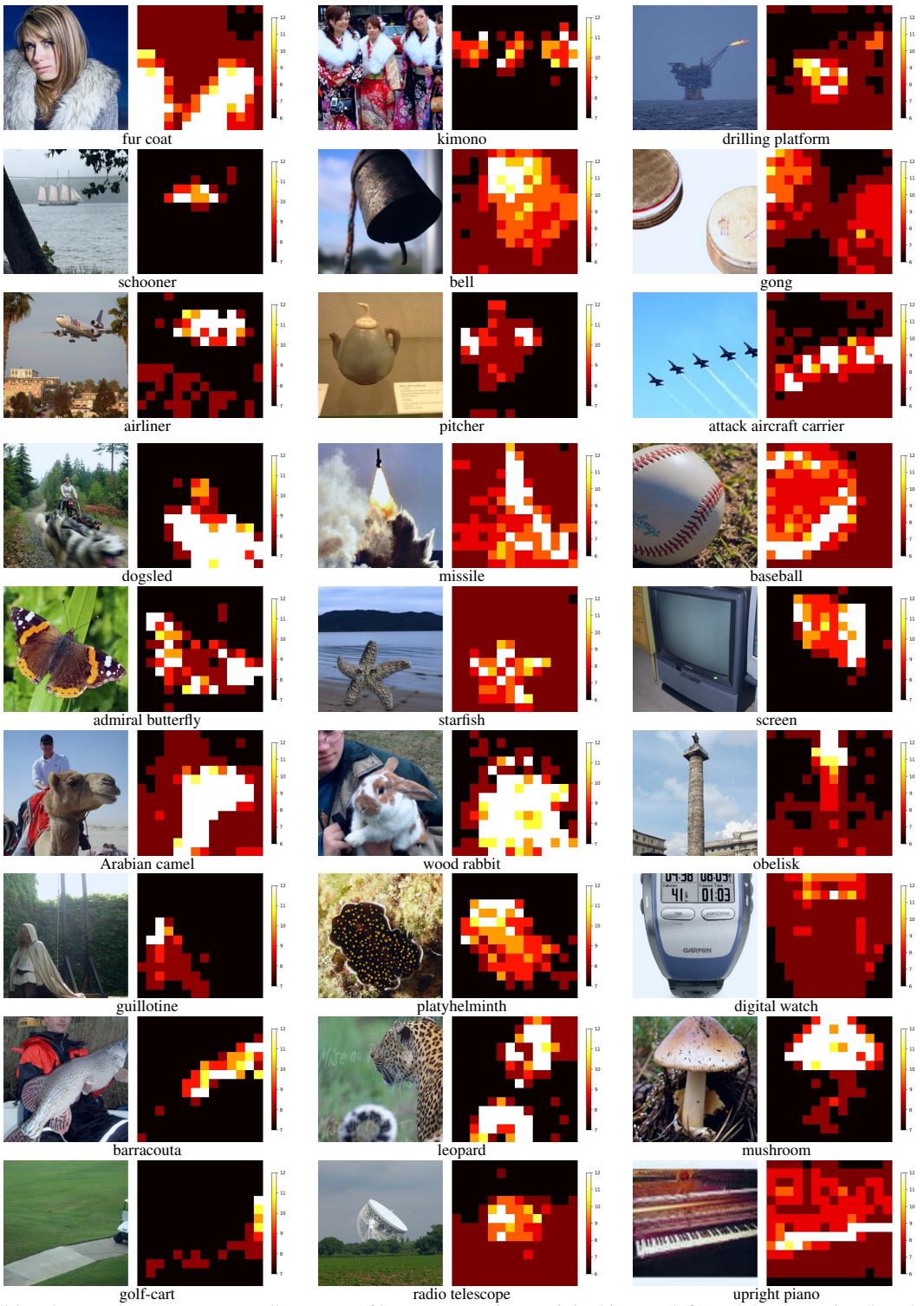

Figure 8. Additional examples across a more diverse set of image categories – original image (left) and the dynamic token depth (right) of A-ViT-T on the ImageNet-1K validation set. Again adaptive tokens can quickly cater to informative regions while filtering out complex backgrounds, *e.g.*, completely ignoring human faces and focusing on the coats, see the first two image on the first row. Even for a very small informative region of the target object, the computation can still be effectively allocated towards it, see the first golf-cart class sample of the last row as an example.

Appendix B - Additional Details

Training. For training setup other than the scaling constants, `lr` specified in the main manuscript, we follow original repository for all other hyper-parameters at https://github.com/facebookresearch/DeiT such as drop out rate, momentum, preprocessing, etc, imposing minimum training recipe changes when adapting a static model to its adaptive counterpart.

Latency. We measure the latency on an NVIDIA TITAN RTX 2080 GPU with PyTorch for batch size of 64 images, CUDA 10.2. For GPU warming up, 100 forward passes are conducted, and then the median speed of the 1K measurements of the full model latency are reported. The exact same setup is shared across all baseline and proposed methods for a fair comparison.

SOTA baselines.. We followed DeiT's [43] repository[2] for recipes and checkpoints as a common starting point for all experiments. For DynamicViT [36], we used the public repository and script from the authors. For other dynamic approaches from CNN/NLP literature, we re-implemented the methods on DeiT to examine ACT [17] for layer-wise halting, confidence threshold [31] on post-softmax logits, two variants of similarity gauging [12] on delta-logits based on (i) LPIPS and (ii) MSE similarity scores, and PonderNet [1] with geometric-distribution sampling towards token halting. For all methods, a detailed grid search was conducted to ensure optimal hyper-parameters.

[2] https://github.com/facebookresearch/DeiT