Supplementary Material for “Selective Kernel Networks”

A. Details of the Compared Models in Table 3

We provide more details for the variants of ResNeXt-50 in Table 3 in the main body of the paper. Compared with this baseline, “ResNeXt-50, wider” has $\frac{1}{16}$ more channels in all bottleneck blocks; “ResNeXt-56, deeper” has extra 2 blocks in the end of the fourth stage of ResNeXt-50; “ResNeXt-50 (36×4d)” has a cardinality of 36 instead of 32. These three structures match the overall complexity of SKNet-50, which makes the comparisons fair.

B. Details of the Models in Table 4

For fair comparisons, we re-implement the ShuffleNetV2 [6] with 0.5× and 1.0× settings (see [6] for details). Our implementation changes the numbers of blocks in the three stages from {4,8,4} to {4,6,6}, therefore the performances and computational costs are slightly different from those reported in the original paper [6] (see Table 4 in the main body of the paper for detailed results). In Table 4, “+ SE” means that the SE module [4] is integrated after each shuffle layer in ShuffleNetV2, “+ SK” means that each 3×3 depthwise convolution is replaced by a SK unit with $M = 2$ (K3 and K5 kernels are used in the two paths, respectively), $r = 4$ and G is the same as the number of channels in the corresponding stage due to the depthwise convolution.

Note that the 3×3 depthwise convolution in the original ShuffleNet is not followed by a ReLU activation function. We verify that the best practice for integrating SK units into ShuffleNet is also without ReLU activation functions in both paths in each SK unit (Table S1).

K3 + ReLU ?	K5 + ReLU ?	Top-1 error (%)
✓	✗	28.65
✗	✓	28.40
✗	✗	**28.36**
✓	✓	28.49

Table S1. Influence of activation functions in two paths of SK units based on ShuffleNetV2 1.0×. Single 224×224 crop is used for evaluation on the ImageNet validation set.

C. Details of the Compared Models in Figure 2

We have plotted the results of some state-of-the-art models including ResNet, ResNeXt, DenseNet, DPN and SENet in Figure 2 in the main body of the paper. Each dot represents a variant of certain model. Table S2 shows the settings of these variants, the numbers of parameters, and the evaluation results on the ImageNet validation set. Note that SENets are based on the corresponding ResNeXts.

Method	#P	Top-1 error (%)
ResNet-50 [3]	25.56M	23.9
ResNet-101 [3]	44.55M	22.6
ResNet-152 [3]	60.19M	21.7
DenseNet-169 (k=32) [5]	14.15M	23.8
DenseNet-201 (k=32) [5]	20.01M	22.6
DenseNet-264 (k=32) [5]	33.34M	22.2
DenseNet-232 (k=48) [5]	55.80M	21.3
ResNeXt-50 (32×4d) [7]	25.00M	22.2
ResNeXt-101 (32×4d) [7]	44.30M	21.2
DPN-68 (32×4d) [1]	12.61M	23.7
DPN-92 (32×3d) [1]	37.67M	20.7
DPN-98 (32×4d) [1]	61.57M	20.2
SENet-50 [4]	27.7M	21.12
SENet-101 [4]	49.2M	20.58
SKNet-26	16.8M	22.74
SKNet-50	27.5M	20.79
SKNet-101	48.9M	20.19

Table S2. The top-1 error rates (%) on the ImageNet validation set with single 224×224 crop testing.

D. Implementation Details on CIFAR Datasets (Section 4.2)

On CIFAR-10 and CIFAR-100 datasets, all networks are trained on 2 GPUs with a mini-batch size 128 for 300 epochs. The initial learning rate is 0.1 for CIFAR-10 and 0.05 for CIFAR-100, and is divided by 10 at 50% and 75% of the total number of training epochs. Following [3], we

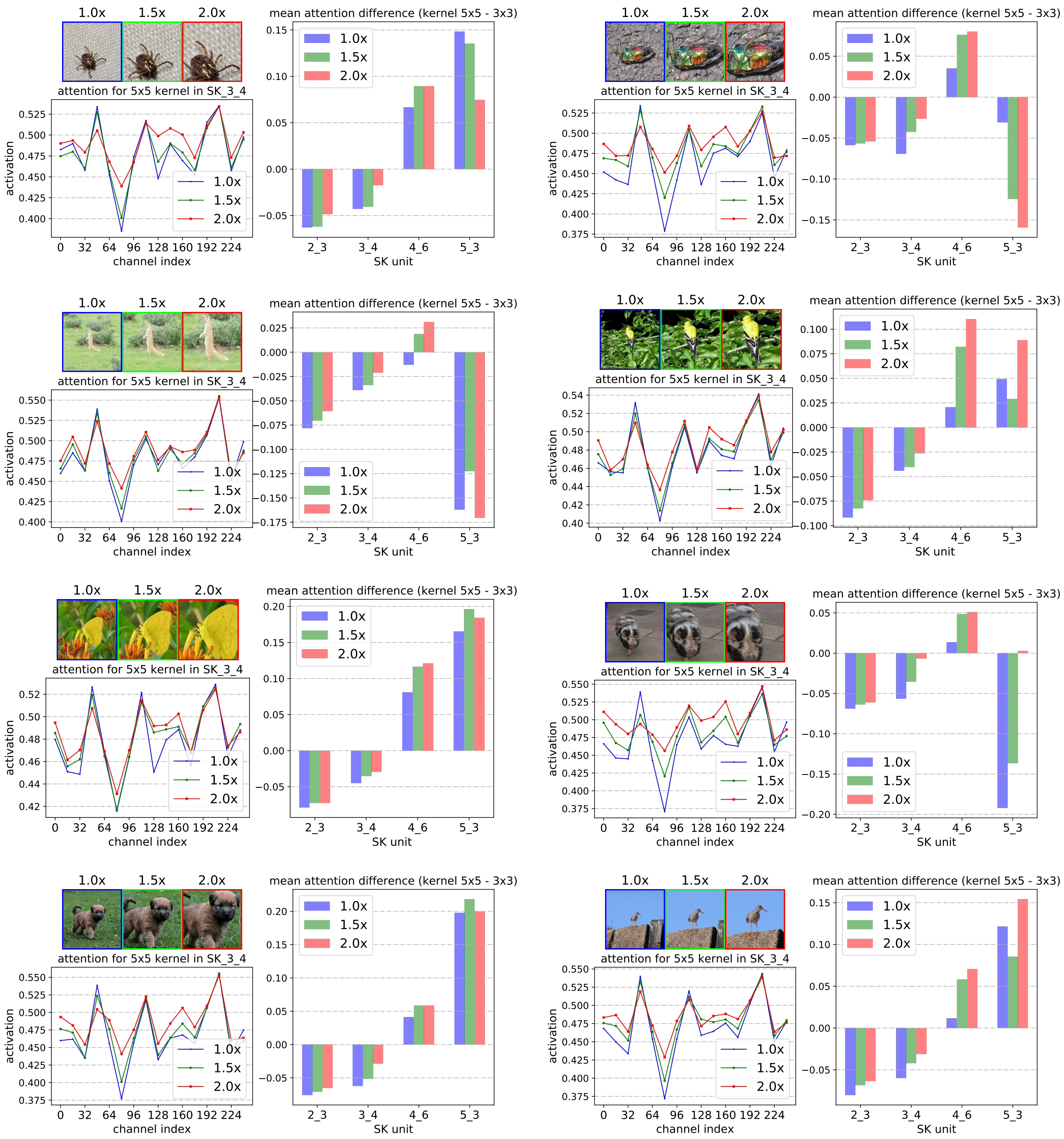

Figure S1. Attention results for two randomly sampled images with three differently sized targets (1.0x, 1.5x and 2.0x). The notations are the same as in Figure 3a,b.

use a weight decay of 5e-4 and a momentum of 0.9. We adopt the weight initialization method introduced in [2]. The ResNeXt-29 backbone is described in [7]. Based on it, SENet-29 applies SE unit before each residual connection, and SKNet-29 modifies the grouped 3×3 convolution to SK convolution with setting SK[2, 16, 32]. In order to prevent overfitting on these smaller datasets, we replace the 5×5 kernel in the second path in the SK unit to 1×1, while the setting for the first path remains the same.

E. More Examples of Dynamic Selection

Figure S1 shows attention results for more images with three differently sized targets. Same as in Figure 3 in the main body of the paper, we see a trend in low and middle level stages: the larger the target object is, the more attention is assigned to larger kernels by the dynamic selection mechanism.

References

[1] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. Dual path networks. In *NIPS*, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017.

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[6] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. *arXiv preprint arXiv:1807.11164*, 2018.

[7] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.