Supplementary Material: Omni-Scale Feature Learning for Person Re-Identification

Kaiyang Zhou[1]* Yongxin Yang[1] Andrea Cavallaro[2] Tao Xiang[1,3]
[1]University of Surrey [2]Queen Mary University of London
[3]Samsung AI Center, Cambridge
{k.zhou, yongxin.yang, t.xiang}@surrey.ac.uk a.cavallaro@qmul.ac.uk

A. More Visualisation on Person ReID

In addition to Fig. 6 in the main paper, Fig. A here provides more examples of activation maps to support the claim that OSNet can learn discriminative features with homogeneous and heterogeneous scales. It can be observed from Fig. A that OSNet is able to identify local patterns with their context as the focus of attention. For example, Figs. A(d) and (f) show that both the T-shirts and the logos are selected for feature extraction. In contrast, the single-scale baseline tends to only focus on local regions while ignoring the contextual information. This renfers the model more susceptible to occlusion and ambiguity of small local patterns.

B. Implementation Details on PA-100K

A sigmoid-activated attribute prediction layer is added on the top of OSNet. Following [4, 5], we use the weighted multi-label classification loss for supervision. For data augmentation, we adopt random translation and mirroring. OSNet is trained from scratch with SGD, momentum of 0.9 and initial learning rate of 0.065 for 50 epochs. The learning rate is decayed by 0.1 at 30 and 40 epochs.

C. Evaluation on ImageNet

In Sec. 4.3 of the main paper, we have reported results of OSNet on the object category recognition tasks of CIFAR10/100. In this section, the results on the larger-scale ImageNet 1K category dataset (LSVRC-2012 [1]) are discussed.

Implementation OSNet is trained with SGD, initial learning rate of 0.4, batch size of 1024 and weight decay of 4e-5 for 120 epochs. For data augmentation, we use random 224 × 224 crops on 256 × 256 images and random mirroring. To benchmark, we report single-crop[1] top1 accuracy on the LSVRC-2012 validation set [1].

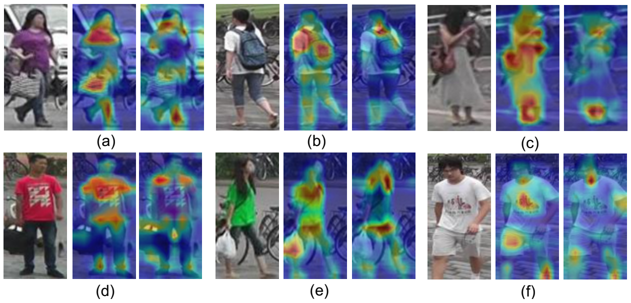

Figure A. Visualisation of activation maps obtained by OSNet (middle one in each triplet) and the single-scale baseline (right one in each triplet).

Method	β	# params	Mult-Adds	Top1
SqueezeNet [3]	1.0	1.2M	-	57.5
MobileNetV1 [2]	0.5	1.3M	149M	63.7
MobileNetV1 [2]	0.75	2.6M	325M	68.4
MobileNetV1 [2]	1.0	4.2M	569M	70.6
ShuffleNet [7]	1.0	2.4M	140M	67.6
ShuffleNet [7]	1.5	3.4M	292M	71.5
ShuffleNet [7]	2.0	5.4M	524M	73.7
MobileNetV2 [6]	1.0	3.4M	300M	72.0
MobileNetV2 [6]	1.4	6.9M	585M	74.7
OSNet (ours)	0.5	1.1M	424M	69.5
OSNet (ours)	0.75	1.8M	885M	73.5
OSNet (ours)	1.0	2.7M	1511M	**75.5**

Table A. Single-crop top1 accuracy (%) on ImageNet-2012 validation set. β: width multiplier. M: Million.

Results Table A shows that OSNet outperforms the alternative lightweight models by a clear margin. In particular OSNet×1.0 surpasses MobiltNetV2×1.0 by 3.5% and MobiltNetV2×1.4 by 0.8%. It is noteworthy that MobiltNetV2×1.4 is around 2.5× larger than our OSNet×1.0. OSNet×0.75 performs on par with ShuffleNet×2.0 and outperforms ShuffleNet×1.5/×1.0 by 2.0%/5.9%. These results give a strong indication that OSNet has a great potential for a broad range of visual recognition tasks. Note that although the model size is smaller, our OSNet does have a higher number of mult-adds operations than its main competitors. This is mainly due to the multi-stream design. However, if both model size and number of

*Work done as an intern at Samsung AI Center, Cambridge.

[1] 224 × 224 centre crop from 256 × 256.

Multi-Adds need to be small for a certain application, we can reduce the latter by introducing pointwise convolutions with group convolutions and channel shuffling [7].

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[2] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1

[3] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 1

[4] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, 2015. 1

[5] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017. 1

[6] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 1

[7] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 2018. 1, 2