Supplementary Material for "Distilling Knowledge From a Deep Pose Regressor Network"

Muhamad Risqi U. Saputra, Pedro P. B. de Gusmao, Yasin Almalioglu, Andrew Markham, Niki Trigoni
Department of Computer Science, University of Oxford
firstname.lastname@cs.ox.ac.uk

Table 1. Comparison with other approaches for $d_{rate} = 65.77\%$

Method	RMS RPE (**t**)	RMS RPE (**r**)	RMS ATE
Supervised T	0.1197	0.2377	26.7386
Supervised S	0.1499	0.1187	65.4179
KD [3]	0.2979	0.1468	120.4039
Chen's OD [1]	0.1567	0.1560	40.6333
FitNets [4]	0.1031	0.1434	28.5408
Ours	**0.1023**	**0.1262**	**18.0915**

1. Comparison with Related Works for $d_{rate} = 65.77\%$

In the main paper Section 7.5, we compare our proposed approach with other distillation approaches for classification and object detection using distillation rate $d_{rate} = 92.95\%$. We use $d_{rate} = 92.95\%$ to show that even in extreme conditions, our proposed approach still yields reasonable performances for Visual Odometry (VO), and better than the competing approaches. However, we also perform experiment with $d_{rate} = 65.77\%$ to compare the accuracy with other approaches when they are expected to maximally perform (it can be seen from Figure 6 in the main paper that our proposed approach reaches the maximum performance with $d_{rate} = 65.77\%$). Table 1 shows the result of this experiment. It can be seen that our proposed approach yields the best performance both in terms of RPE or ATE although the competing approaches also show improvement compared to the result from Table 3 in the main paper.

2. Training Time and Convergence

The training time for our approach takes around 3.5 hours for each stage in TITAN V (2x faster than training T). The training time is relatively the same for different d_{rate} as we initialize S with FlowNet [2] and only train the remaining layers using our proposed approach. Fig. 1 shows the validation loss during training between supervised T and distilled S for the second stage of training. Note that we can only show the validation loss for the second stage of training as the first stage of training is intended to learn the intermediate representation which is not comparable with the validation loss for the final task. Note also that T was trained for maximum 200 epochs or was stopped earlier if validation loss showed no improvement, while S was only trained for 30 epochs for each stage. Fig. 1 shows that S converges faster than T as the knowledge is transferred effectively from T.

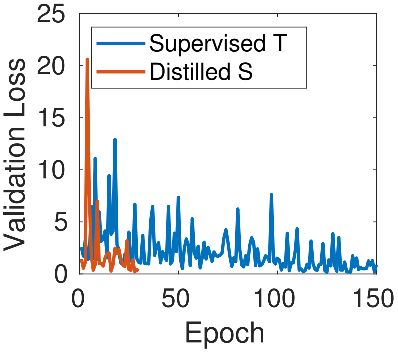

Figure 1. Training convergence between supervised T and distilled S (for the second stage).

References

[1] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 742–751, 2017. 1

[2] A. Dosovitskiy, P. Fischery, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox. FlowNet: Learning Optical Flow with Convolutional Networks. In *IEEE International Conference on Computer Vision (ICCV)*, volume 11-18-Dece, pages 2758–2766, 2016. 1

[3] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop (2015)*, pages 1–9, 2015. 1

[4] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. FitNets: Hints for Thin Deep Nets. In *International Conference on Learning Representations (ICLR)*, 2015. 1