Supplementary: Self-distillation and Uncertainty Boosting Self-supervised Monocular Depth Estimation

Hang Zhou
hang.zhou@uea.ac.uk
David Greenwood
david.greenwood@uea.ac.uk
Sarah Taylor
s.l.taylor@uea.ac.uk
Michal Mackiewicz
M.Mackiewicz@uea.ac.uk

School of Computing Sciences
University of East Anglia
Norwich, UK

In Table 1, in addition to our final uncertainty modelling scheme (last row), we experiment with two input settings for proposed Photometric UncertNet (first two rows) and a standalone uncertainty network (3rd row) for self-distillation uncertainty modelling.

Table 1: SUB-Depth experiments with different uncertainty settings. 1: feeding first frame into UncertNet. 2: feeding first frame and aligned frame into UncerNet. 3: standalone regression uncertainty network.

idx	Abs Rel	Sq Rel	RMSE log	δ_1	δ_2	δ_3
1	0.113	0.905	0.189	0.882	0.961	0.982
2	0.111	0.875	0.188	0.882	0.960	0.982
3	0.111	0.870	0.188	0.883	0.961	0.982
Final	**0.110**	**0.821**	**0.185**	**0.884**	**0.962**	**0.983**

In Table 2, we extended our quantitative evaluation by selecting the top 10 most challenging images for each model, following the method described by [6]. The top 10 hardest images show areas of deep shadow, poor lighting, foliage and other photographically indistinct regions. Our method deals with this uncertainty and improves on the results of all prior methods for this subset of the benchmark test set.

Qualitative evaluations are provided in Figure 1 for randomly selected examples. For each example, we show input RGB and output depth and regression uncertainty maps. The uncertainty map correctly marks object boundaries with high values where the transition from near to far distance is more difficult to predict. To show generalisation performance, in Figure 2, we additionally qualitatively evaluate the same depth network on the Cityscapes dataset [1]. Although trained only on KITTI, the model appears to generalise well for both depth estimation and uncertainty modelling.

As KITTI does not have dense ground truth depth maps, we use Virtual KITTI [2] to compute depth error maps in Figure 3. In this qualitative evaluation we show, from top to bottom, the input RGB image, the depth error maps from the baseline Monodepth2 model

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

Table 2: **Quantitative comparison of SUB-Depth to existing SDE framework trained models on top-10 selected subset of KITTI [3] benchmark**. The best results in each subsection are in **bold**. Models trained with SUB-Depth outperform the same models trained with SDE in every case.

Method	Abs Rel	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
Monodepth2 [4]	0.250	3.008	7.515	0.353	0.683	0.870	0.924
+ SUB-Depth	**0.229**	**2.451**	**6.885**	**0.330**	**0.713**	**0.876**	**0.931**
Improvement	0.021	0.557	0.63	0.023	0.030	0.006	0.007
HR-depth [5]	0.240	1.687	5.433	0.320	0.669	0.871	0.947
+ SUB-Depth	**0.222**	**1.566**	**5.176**	**0.304**	**0.710**	**0.891**	**0.949**
Improvement	0.018	0.121	0.257	0.016	0.041	0.020	0.002
DIFFNet [6]	0.225	2.160	6.357	0.312	0.712	0.899	0.951
+ SUB-Depth	**0.209**	**1.672**	**5.783**	**0.294**	**0.723**	**0.907**	**0.957**
Improvement	0.016	0.488	0.574	0.018	0.011	0.008	0.006

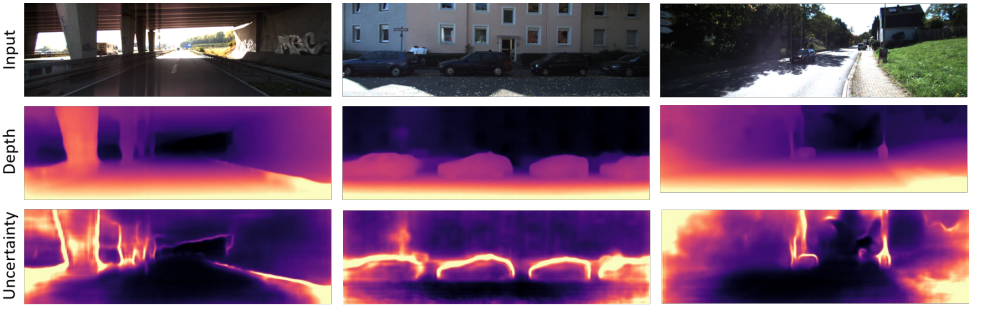

Figure 1: **Qualitative results on KITTI [3]**. We visualise the depth and the uncertainty maps from SUB-Depth trained Monodepth2. The uncertainty maps capture high uncertainty at object boundaries with a hotter color.

and the error maps from Monodepth2 trained with SUB-Depth. For each randomly selected example, we highlight regions of the depth maps that show compelling improvements over prior work. The images are provided at high resolution to allow the reader to zoom in.

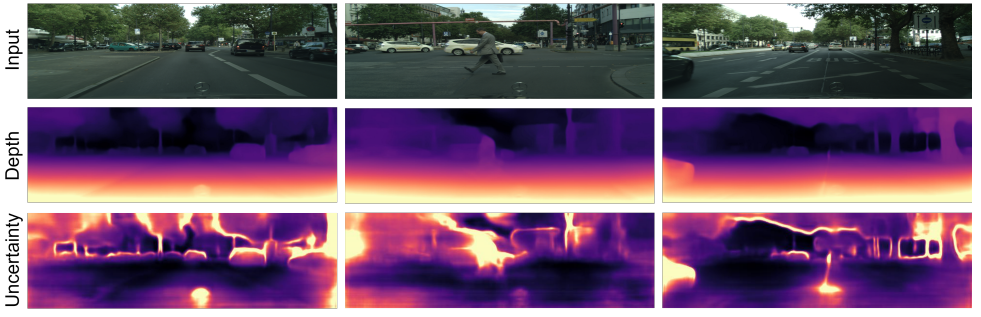

Figure 2: **Generalisation results on Cityscapes [1].** We visualise the depth and the uncertainty maps from SUB-Depth trained only with KITTI. The uncertainty maps show higher uncertainty with a hotter color, and illustrate greater uncertainty at object boundaries and for moving objects.

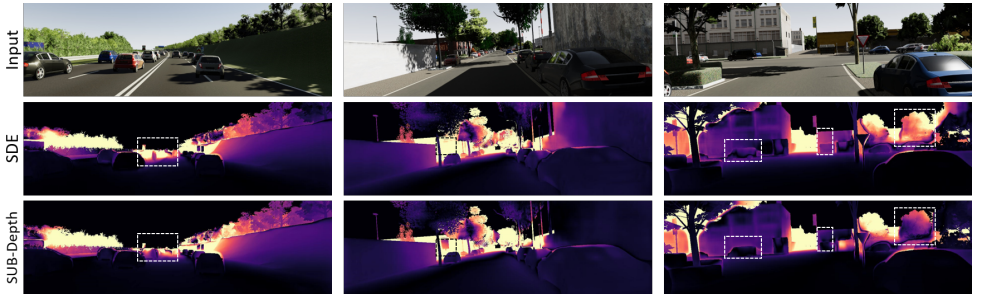

Figure 3: **Visualisation of error map on Virtual KITTI [2].** The top row contains the synthetic input images. The second row shows the Abs rel error maps from SDE trained Monodepth2. The bottom row shows the error maps from SUB-Depth trained Monodepth2. The differences are highlighted by white dotted boxes.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[2] A Gaidon, Q Wang, Y Cabon, and E Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, 2013.

[4] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. In *International Conference on Computer Vision (ICCV)*, 2019.

[5] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. *In AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[6] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. In *British Machine Vision Conference (BMVC)*, 2021.