Supplementary Material

A. KITTI Improved Ground Truth

The evaluation method that was introduced by Eigen *et al*. [2] uses reprojected LiDAR points to create the ground truth depth images. However, the reprojections do not handle occlusions, non-rigid motion or motion from the camera. Uhrig *et al*. [12] introduced an improved set of high quality ground truth depth maps for the KITTI dataset. These high quality images are instead reprojected using 5 consecutive LiDAR frames and uses the stereo images for better handling of occlusions. To obviate the need of retraining, as with other work [5], we use a modified Eigen [2] test split on the images that overlap between these datasets. This results in 652 (93%) of the 697 original test frames being retained. We use the same evaluation strategy and metrics as discussed in the Experiments section of the main paper. The results of this analysis can be found in Table 2.

B. Network Architecture

For all experiments, except where noted, we use a ResNet-101 encoder model with pretrained ImageNet weights. This model has been modified to use atrous/dilation convolutions [1] in the final two residual blocks. We use rectified linear activation (ReLU) in the encoding model and the Exponential Linear Unit (ELU) in the decoder. Skip connections are applied to the two intermediate outputs between the encoder and decoder. As the internal resolution is much larger than that of the ResNet-18 used by Monodepth2 [5] ($\frac{1}{8}$ scale compared with $\frac{1}{32}$ scale), a skip connection is not required for the smallest output resolution. For the pose model, we use the same ResNet-18 and pose decoder defined by Monodepth2 [5]. The full depth network architecture can be found in Table 1.

Depth Network (Number of Parameters: 51.34M)							
layer	**k**	**s**	**ch**	**dilation**	**res**	**input**	**activation**
conv1	3	1	64	2	1	image	ReLU
conv2	3	1	64	1	2	conv1	ReLU
conv3	3	1	128	1	2	conv2	ReLU
maxpool	3	2	128	1	2	conv2	ReLU
res1	3	1	256	1	4	conv3	ReLU
res2	3	2	512	1	8	res1	ReLU
res3	3	1	1024	2	8	res2	ReLU
res4	3	1	2048	4	8	res4	ReLU
context	3	1	512	1	8	res4	Self-Attn
ddv4	3	1	128	1	8	context	Linear
disp4	3	1	1	1	8	ddv1	softmax
upconv3	3	1	64	1	8	ddv4	ELU
deconv3	3	1	64	1	4	upconv3↑, res1	ELU
ddv3	3	1	128	1	4	deconv3	Linear
disp3	3	1	1	1	4	ddv3	softmax
upconv2	3	1	64	1	4	deconv3	ELU
deconv2	3	1	64	1	2	upconv2↑, conv3	ELU
ddv2	3	1	128	1	2	deconv2	Linear
disp2	3	1	1	1	2	ddv2	softmax
upconv1	3	1	32	1	2	deconv2	ELU
deconv1	3	1	32	1	1	upconv1↑	ELU
ddv1	3	1	128	1	1	deconv1	Linear
disp1	3	1	1	1	1	ddv1	softmax

Table 1. **Network architecture**. This table details the kernel size (**k**), stride (**s**), output channels (**ch**) dilation factor (**dilation**), resolution scale (**res**), input features for each layer (**input**) and activation function (**activation**) used in our model. Layers marked with ↑ represent a $2\times$ nearest-neighbour upsampling before passing to the convolutional layer. Residual blocks are denoted by $res*$ naming convention. Each convolution and residual block also uses batch normalisation in the form of a inplace activated batch normalisation [11]. The self-attention module (*context*) is denoted as having an activation of *Self-Attn*.

C. Additional Qualitative Results

In Figure 1, we present additional qualitative comparisons to multiple previous works. Our method produces sharper predictions for thin structures and complex shapes such as people. In Figure 2, we show the uncertainty estimates for multiple images. As can been seen in the figure, areas of low contrast (row 2) correspond with areas of high uncertainty. Moreover, high uncertainty can also be observed in areas of unknown texture (row 7, right hand side). This area of the input image also demonstrates issues with texture copy artefacts [5] in the predicted depth. Additional attention maps are displayed in Figure 3. The attention maps were selected at random from the 512 output channels in the context module.

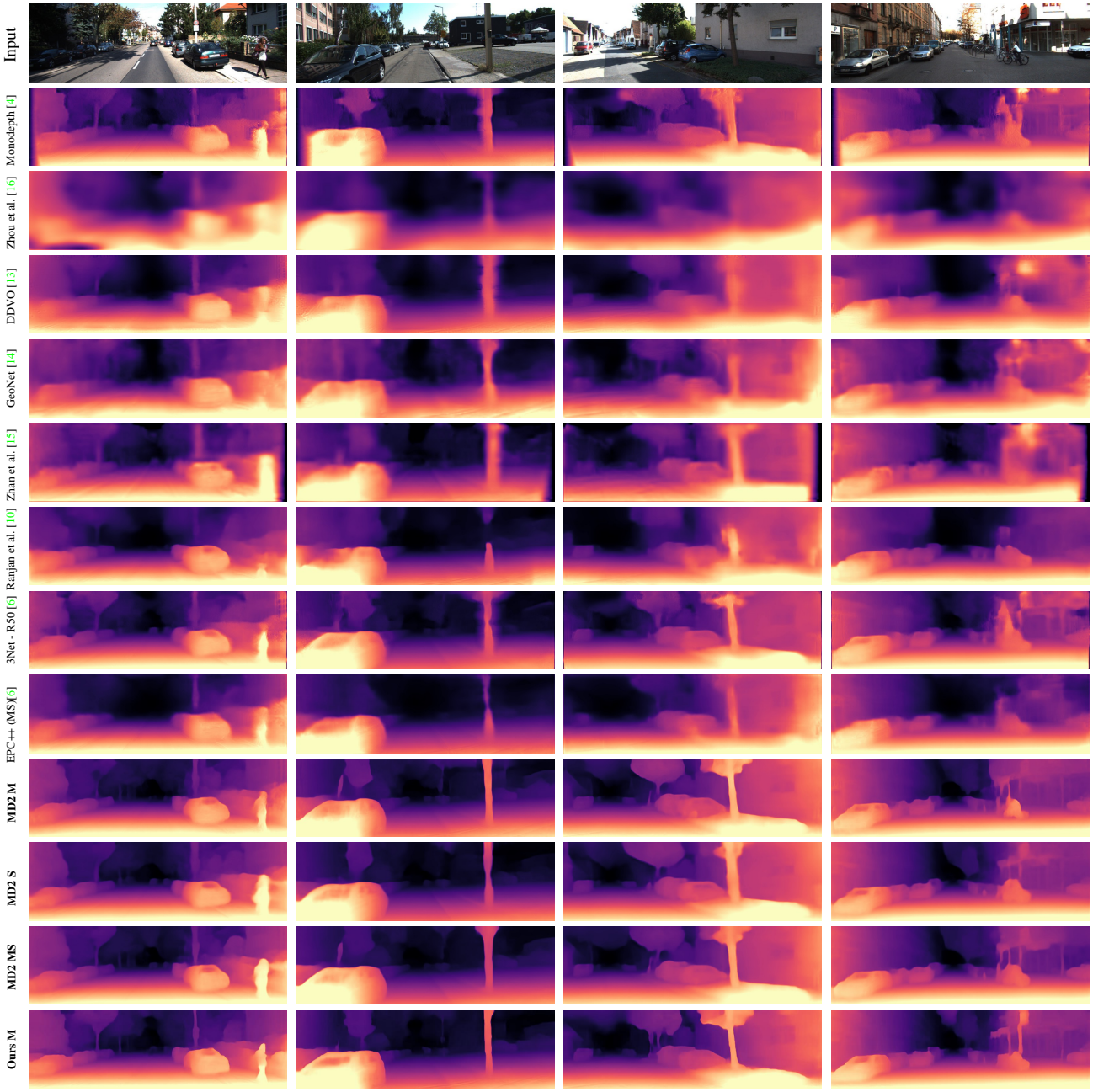

Figure 1. **Additional Qualitative Comparison.** A comparison of our method (*last row*) with several other methods for monocular and stereo trained self supervised depth estimation.

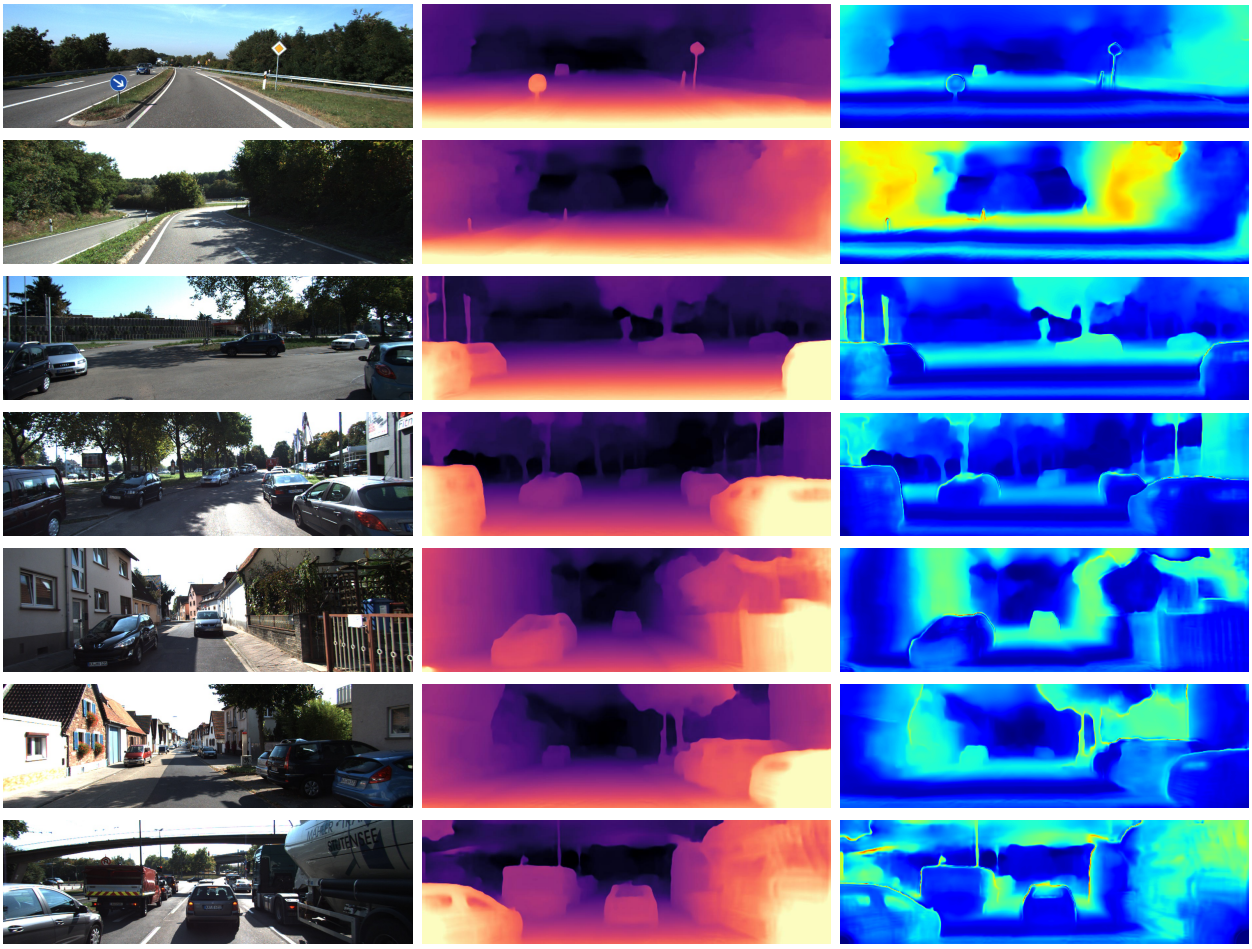

Figure 2. **Additional uncertainty results** The Discrete Disparity Volume (DDV) allows us to compute pixel-wise depth uncertainty by measuring the variance across the disparity *ray*. Left: Input Image, Middle: Depth prediction, Right: Uncertainty (Blue indicates areas of low uncertainty, green/red regions indicate areas of high/highest uncertainty).

Figure 3. **Additional attention maps** selected at random from the output of context module (Blue indicates areas of high attention).

Method	Train	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou [16]†	M	0.176	1.532	6.129	0.244	0.758	0.921	0.971
Mahjourian [7]	M	0.134	0.983	5.501	0.203	0.827	0.944	0.981
GeoNet [14]	M	0.132	0.994	5.240	0.193	0.833	0.953	0.985
DDVO [13]	M	0.126	0.866	4.932	0.185	0.851	0.958	0.986
Ranjan [10]	M	0.123	0.881	4.834	0.181	0.860	0.959	0.985
EPC++ [6]	M	0.120	0.789	4.755	0.177	0.856	0.961	0.987
Monodepth2 [5] w/o pretraining	M	0.112	0.715	4.502	0.167	0.876	0.967	0.990
Monodepth2 [5]	M	0.090	0.545	3.942	0.137	0.914	0.983	0.995
Ours	M	**0.081**	**0.484**	**3.716**	**0.126**	**0.927**	**0.985**	**0.996**
Monodepth [4]	S	0.109	0.811	4.568	0.166	0.877	0.967	0.988
3net [9] (VGG)	S	0.119	0.920	4.824	0.182	0.856	0.957	0.985
3net [9] (ResNet 50)	S	0.102	0.675	4.293	0.159	0.881	0.969	0.991
SuperDepth [8] + pp	S	0.090	0.542	3.967	0.144	0.901	0.976	**0.993**
Monodepth2 [5] w/o pretraining	S	0.110	0.849	4.580	0.173	0.875	0.962	0.986
Monodepth2 [5]	S	**0.085**	**0.537**	**3.868**	**0.139**	**0.912**	**0.979**	**0.993**
Zhan FullNYU [15]	D*MS	0.130	1.520	5.184	0.205	0.859	0.955	0.981
EPC++ [6]	MS	0.123	0.754	4.453	0.172	0.863	0.964	0.989
Monodepth2[5] w/o pretraining	MS	0.107	0.720	4.345	0.161	0.890	0.971	0.989
Monodepth2[5]	MS	**0.080**	**0.466**	**3.681**	**0.127**	**0.926**	**0.985**	**0.995**

Table 2. **Quantitative results on KITTI improved ground truth.** Comparison of existing methods to our own on the KITTI 2015 [3] using the improved ground truth [12] of the Eigen test split [2]. The Best results are presented in **bold** for each category, with second best results underlined. The supervision level for each method is presented in the *Train* column with; D – Depth Supervision, D* – Auxiliary depth supervision, S – Self-supervised stereo supervision, M – Self-supervised mono supervision. Results are presented without any post-processing [4], unless marked with – + pp. If newer results are available on github, these are marked with – †. Non-Standard resolutions are documented along with the method name. Metrics indicated by red: *lower is better*, Metrics indicated by blue: *higher is better*

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[2] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 1, 4

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012. 4

[4] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 2, 4

[5] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. *The International Conference on Computer Vision (ICCV)*, October 2019. 1, 4

[6] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3D holistic understanding. *arXiv*, 2018. 2, 4

[7] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. In *CVPR*, 2018. 4

[8] Sudeep Pillai, Rares Ambrus, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *ICRA*, 2019. 4

[9] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *3DV*, 2018. 4

[10] Anurag Ranjan, Varun Jampani, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, 2019. 2, 4

[11] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5639–5647, 2018. 1

[12] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant CNNs. In *3DV*, 2017. 1, 4

[13] Chaoyang Wang, Jose Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018. 2, 4

[14] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 2, 4

[15] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018. 2, 4

[16] Tinghui Zhou, Matthew Brown, Noah Snavely, and David Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2, 4