Supplementary material

This document provides additional details concerned with CVPR 2019 paper "Learning monocular depth estimation infusing traditional stereo knowledge". We have included the detailed specification of our depth-from-mono architecture monocular Residual Matching (*monoResMatch*), additional quantitative results and more visual depth maps on KITTI and CityScapes datasets.

1. Specification of monoResMatch

The detailed specification of our network provided in Table 1 can be divided into three modules: the multi-scale feature extractor, the initial disparity estimator and the final disparity refinement stage. For each layer of the network, we report convolution kernel size **K**, stride **S**, the input and output number of channels and the input of the layer. The symbol "," means concatenation.

2. Depth Estimation with 50 cap

In this section, we report additional experimental results on the Eigen's KITTI test split [1], evaluating depth maps within 0-50 m range. Table 2 shows a comparison between our architecture *monoResMatch* and other works reporting this quantitative evaluation, confirming once again the superiority of our proposal compared to all competitors.

3. Qualitative results

Our paper has shown several disparity maps predicted by the proposed architecture. As a supplement, we report in this document more outcomes of our network using both Cityscapes and KITTI datasets. Figure 1 shows qualitative results on the test set of KITTI stereo 2015 [6] generated by *monoResMatch* fine-tuned on the 200-acrt ground-truth labels of the training set. In Figure 1, the disparity and error images are extracted from the KITTI evaluation website. Disparity images are shown using the color map from [3]. For the error images, warmer color indicates larger errors in depth prediction. In Figure 2, we visualize disparity maps on the Cityscapes dataset. As suggested in [4], we discarded the lower 20% of the input image to delete the hood of the car. Moreover, we resized the input shape to 1024×512 resolution. Figure 4 presents results comparing three different configurations of *monoResMatch*. Specifically, we analyze visual effects of the network trained on 1) Semi-Global Matching proxy labels with those obtained by 2) fine-tuning on 200 accurate ground-truth labels of KITTI 2015 and 3) on 700 raw LiDAR samples from the Eigen training split.

Finally, in Figure 4 we report qualitative comparison with other state-of-the-art methods on the Eigen test split.

Layer	K	S	In/Out	Input
Multi-scale feature extractor				
conv1	7	2	3/64	input
up_conv1	4	2	64/32	conv1
conv2	5	2	64/128	conv1
up_conv2	8	4	128/32	conv2
up_conv12	1	1	64/32	up_conv1, up_conv2
Initial Disparity Estimation				
conv_rdi	1	1	128/64	conv2
conv3	3	2	64/256	conv_rdi
conv3_1	3	1	256/256	conv3
conv4	3	2	256/512	conv3_1
conv4_1	3	1	512/512	conv4
conv5	3	2	512/512	conv4_1
conv5_1	3	1	512/512	conv5
conv6	3	2	512/1024	conv5_1
conv6_1	3	1	1024/1024	conv6
disp6	3	1	1024/2	conv6_1
upconv5	4	2	1024/512	conv6_1
updisp6	4	2	2/1	disp6
iconv5	4	1	1025/512	upconv5,updisp6,conv5_1
disp5	3	1	512/2	iconv5
upconv4	4	2	512/256	iconv5
updisp5	4	2	2/1	disp5
iconv4	4	1	769/512	upconv4,updisp5,conv4_1
disp4	3	1	256/2	iconv4
upconv3	4	2	256/128	iconv4
updisp4	4	2	2/1	disp4
iconv3	4	1	385/128	upconv3,updisp4,conv3_1
disp3	3	1	128/2	iconv3
upconv2	4	2	128/64	iconv3
updisp3	4	2	2/1	disp3
iconv2	4	1	193/64	upconv2,updisp3,conv2_1
disp2	3	1	64/2	iconv2
upconv1	4	2	64/32	iconv2
updisp2	4	2	2/1	disp2
iconv1	4	1	97/32	upconv1,updisp2,conv1_1
disp1	3	1	32/2	iconv1
upconv0	4	2	32/16	iconv1
updisp1	4	2	2/1	disp1
iconv0	4	1	49/32	upconv0,updisp1,up_conv12
disp0	3	1	32/2	iconv0
Warping				
wr_conv1	-	-	64/64	conv1
wr_up_conv12	-	-	32/32	up_conv12
wl_up_conv12	-	-	32/32	wr_up_conv12
Disparity Refinement				
r_conv0	3	1	66/32	|up_conv12 - wl_up_conv12|, disp0, up_conv12
r_conv1	3	2	32/64	r_conv0
c_conv1	3	1	64/16	conv1
wr_c_conv1	3	1	64/16	wr_conv1
r_corr	-	-	16/41	c_conv1, wr_c_conv1
r_conv1_1	3	1	105/64	r_corr, r_conv1
r_conv2	3	2	64/128	r_conv1_1
r_conv2_1	3	1	128/128	r_conv2
r_res2	3	1	130/1	r_conv2_1, disp2
up_r_res2	3	2	1/2	r_res2
r_upconv1	4	2	64/128	r_conv2_1
r_iconv1	4	1	130/64	r_upconv1, up_r_res2, r_conv1_1
r_res1	3	1	66/1	r_iconv1, disp1
up_r_res1	3	2	1/2	r_res1
r_upconv0	4	2	32/64	r_conv1
r_iconv0	4	1	66/32	r_upconv0, up_r_res1, r_conv0
r_res0	3	1	34/1	r_iconv0, disp0

Table 1. Detailed *monoResMatch* architecture.

References

[1] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep net-

Method	Supervision	Train set	Lower is better				Higher is better		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou *et al.* [11]	Seq	CS,K	0.190	1.436	4.975	0.258	0.735	0.915	0.968
Yin *et al.* [10] GeoNet ResNet50	Seq	K	0.147	0.936	4.348	0.218	0.810	0.941	0.977
Zou *et al.* [12]	Seq	CS,K	0.146	1.182	5.215	0.213	0.818	0.943	0.978
Mahjourian *et al.* [5]	Seq	CS,K	0.151	0.949	4.383	0.227	0.802	0.935	0.974
Poggi *et al.* [7] PyD-Net (200)	Stereo	CS,K	0.138	0.937	4.488	0.230	0.815	0.934	0.972
Godard *et al.* [4] ResNet50	Stereo	CS,K	0.108	0.657	3.729	0.194	0.873	0.954	0.979
Poggi *et al.* [8] 3Net ResNet50	Stereo	CS,K	**0.091**	0.572	3.459	0.183	0.889	0.955	0.979
Yang *et al.* [9]	Seq+Stereo	K_o, K_r, K_o	0.092	0.547	3.390	0.177	0.898	0.962	0.982
monoResMatch	Stereo	CS,K	**0.091**	**0.504**	**3.336**	**0.174**	**0.899**	**0.965**	**0.984**

Table 2. Quantitative evaluation on the test set of KITTI dataset [2] using the split of Eigen *et al.* [1], with maximum depth set to 50m. K_o, K_r, K_o are splits from K, defined in [9]. Best results are shown in bold.

work. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 1, 2

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012. 1

[4] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 1, 2, 5

[5] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[6] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[7] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. Towards real-time unsupervised monocular depth estimation on cpu. In *IEEE/JRS Conference on Intelligent Robots and Systems (IROS)*, 2018. 2

[8] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *6th International Conference on 3D Vision (3DV)*, 2018. 2

[9] Nan Yang, Rui Wang, Jörg Stückler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *European Conference on Computer Vision*, pages 835–852. Springer, 2018. 2

[10] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5

[11] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017. 2, 5

[12] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *European Conference on Computer Vision*, pages 38–55. Springer, 2018. 2

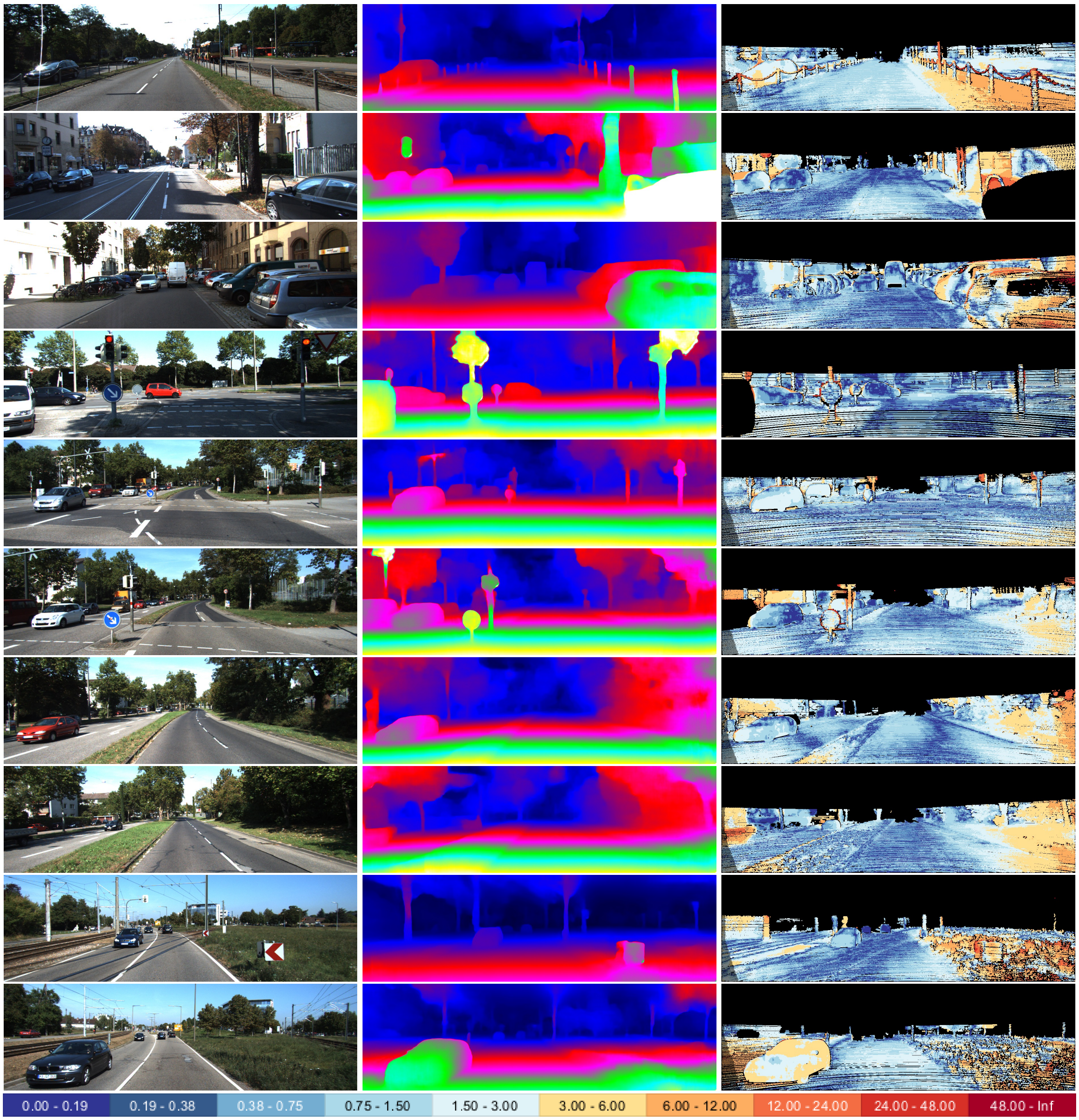


Figure 1. Stereo evaluation of our depth-from-mono framework. From left to right the input image, the predicted disparity and the errors with respect to ground truth. The last line reports the color code used to display the seriousness of the shortcomings.

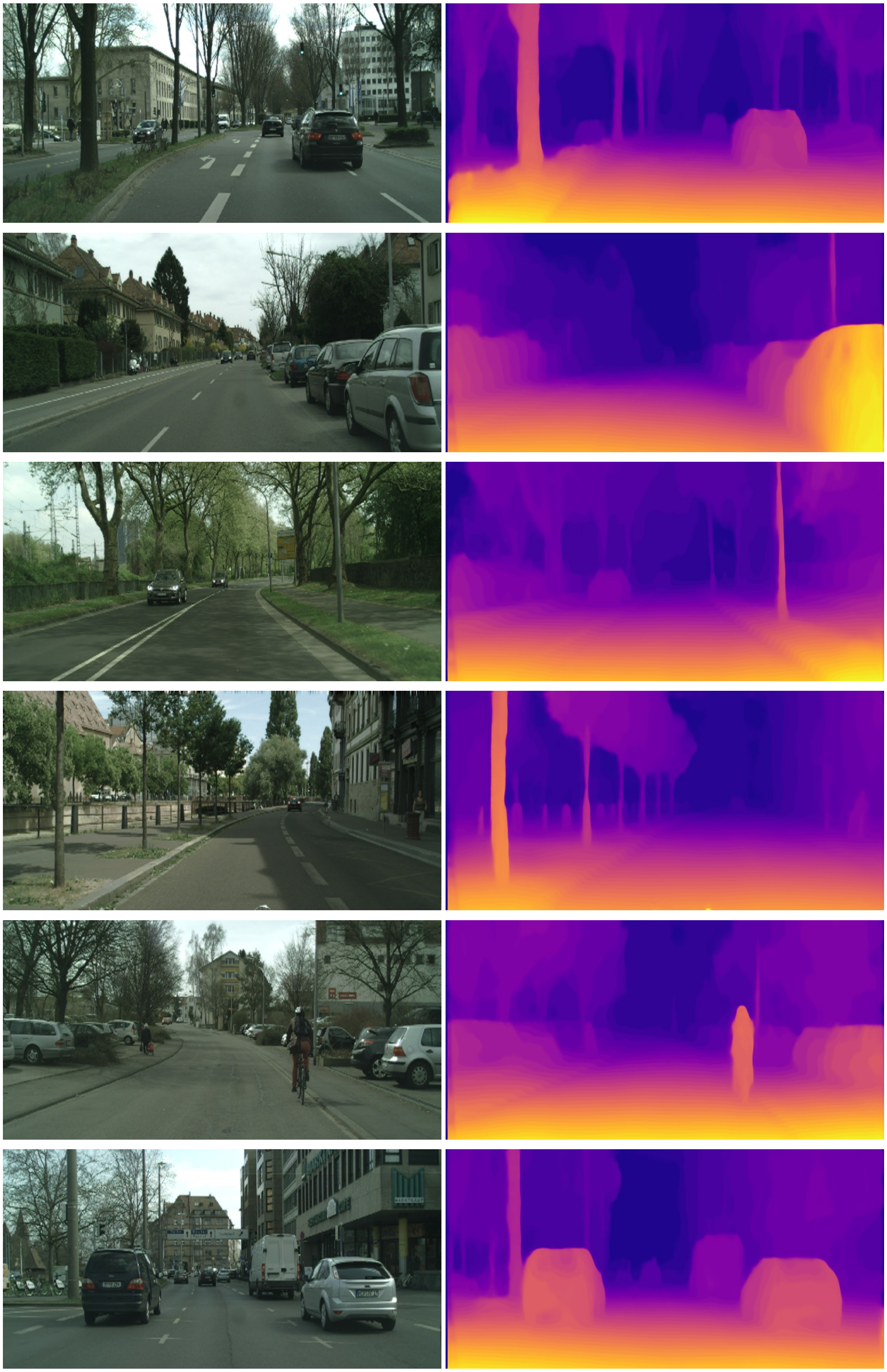
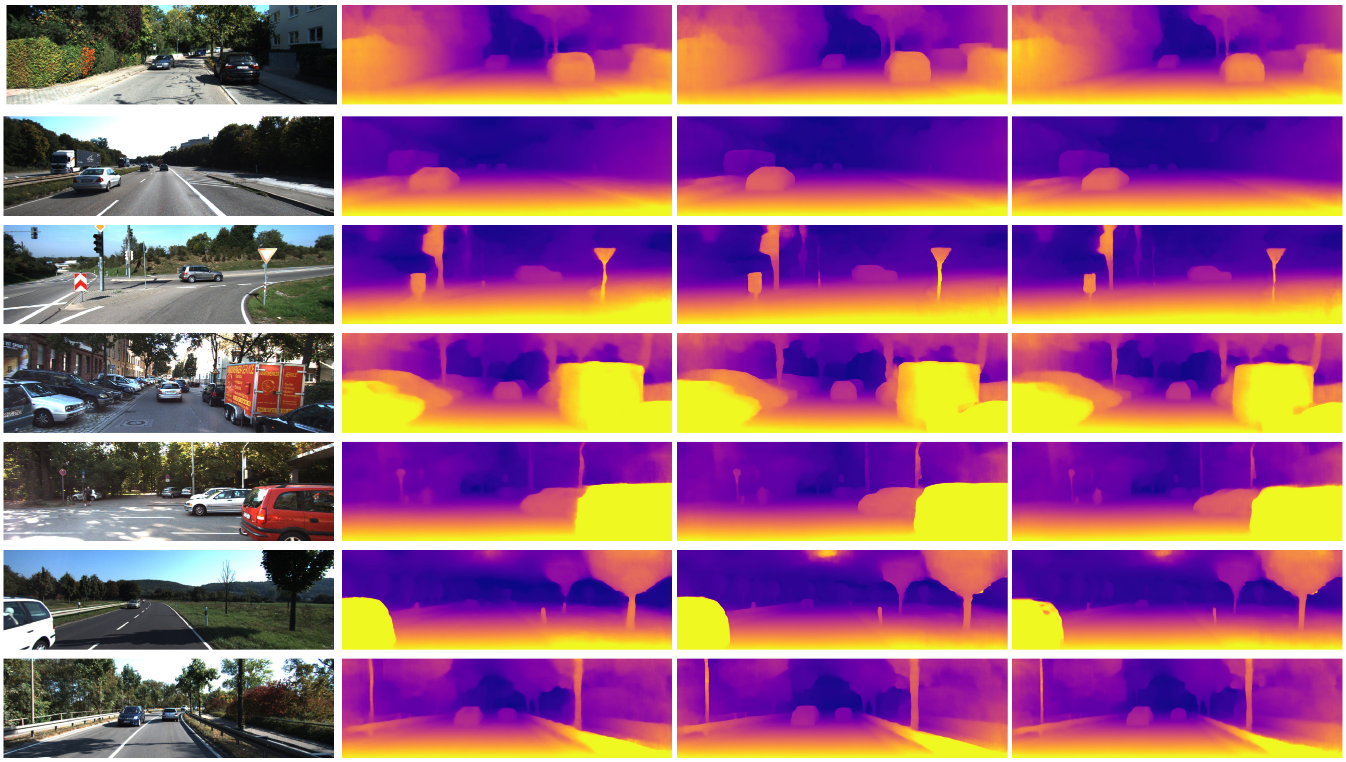
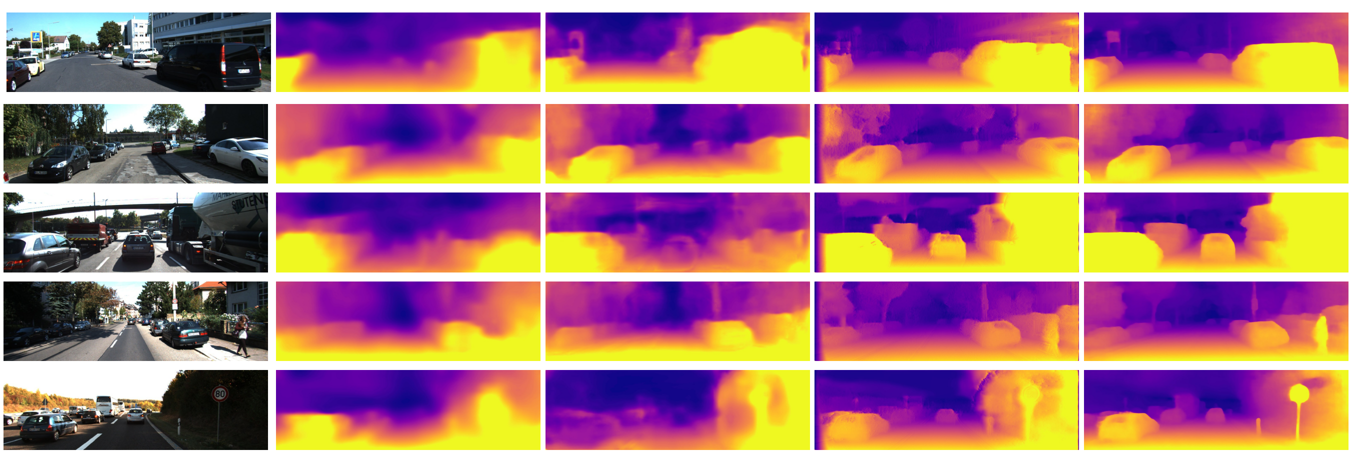


Figure 2. Qualitative results of monoResMatch trained semi-supervisedly using SGM on Cityscapes dataset.



(a) (b) (c) (d)

Figure 3. Qualitative results of the proposed depth-from-mono architecture. From left to right, the input image from KITTI 2015 test set (a), the predicted depth by monoResMatch trained on (b) SGM proxy annotations, fine-tuned using (c) 200-acrt ground-truth labels or (d) 200-acrt + 700 raw LiDAR samples.



Input Image Zhou et al. [11] Yin et al. [10] Godard et al. [4] ours (SGM only)

Figure 4. Qualitative comparison with state-of-the-art methods on Eigen's KITTI test split. For our network and [4] no post-processing operation has been applied.