Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Supplementary Material - On the Benefit of Adversarial Training for Monocular Depth Estimation

Rick Groenendijk[a], Sezer Karaoglu[b], Theo Gevers[a,b], Thomas Mensink[a,c]

[a]University of Amsterdam, Science Park 904, 1098XH Amsterdam, the Netherlands
[b]3DUniversum, Science Park 400, 1098 XH Amsterdam, the Netherlands
[c]Google Research, Claude Debussylaan 34, 1082 MD Amsterdam, the Netherlands

ABSTRACT

In this supplementary report we provide additional experimental results and information on the implementation details belonging to our main paper. These results will be made available in an (online) appendix upon acceptance of the paper.

1. Extended Experimental Results

1.1. Loss Components

In the main work, Eqs. 2,3,5,6 the definitions of $\mathcal{L}_{ap}$, $\mathcal{L}_{lr}$, $\mathcal{L}_{disp}$ were given. In follow-up work Yang et al. (2018) extends the loss function with two components: $\mathcal{L}_{occl}$ and a semi-supervised loss function which takes into account sparse disparity maps for a subset of images in the training set. The semi-supervised component is ignored in this work. The new loss definition is:

$$\mathcal{L}_s = \gamma_{L1}\, \mathcal{L}_{L1} + \gamma_S\, \mathcal{L}_S + \gamma_{lr}\, \mathcal{L}_{lr} + \gamma_{disp}\, \frac{1}{2^s}\mathcal{L}_{disp} + \gamma_{occl}\, \mathcal{L}_{occl}, \tag{1}$$

where $\mathcal{L}_{occl}$ is the occlusion loss. The occlusion loss penalizes the total sum of disparities, to favor background depths. Also combining the occlusion loss with the disparity gradient loss enforces transitions at occlusions. These occlusions happen due to the stereo set-up of the cameras.

$$\mathcal{L}^l_{occl} = \frac{1}{\mathrm{N}} \sum_{i,j} |d^l_{ij}| \tag{2}$$

In the supplementary material of Yang et al. (2018) it is suggested that disparity smoothness loss $\mathcal{L}_{disp}$ in itself does not improve model performance. However a combination of $\mathcal{L}_{disp}$ and $\mathcal{L}_{occl}$ does seem to increase model performance. The parameters that are used in the papers are as follows:

- Godard et al. (2017): $\gamma_{l1} = 0.15$, $\gamma_{l1} = 0.85$, $\gamma_{lr} = 1.0$, $\gamma_{disp} = 0.1$
- Yang et al. (2018): $\gamma_{l1} = 0.15$, $\gamma_{l1} = 0.85$, $\gamma_{lr} = 1.0$, $\gamma_{disp} = 0.1$, $\gamma_{occl} = 0.01$

An ablation study of loss components is conducted. 2 shows the results. Unlike Yang et al. (2018) no significant quantitative benefit is found of using the occlusion loss component. Moreover, it seems state-of-the-art performance can be acquired through only using SSIM loss as the only loss component.

We also detail the main experiment for even more settings than shown in the paper. We evaluate more settings of loss combinations with and without adversarial training. The results are shown in Tab. 1. From the table, the main results of the paper are concluded. Furthermore it becomes clear that adding the LR loss yields a large performance boost (experiment #2 vs #3), and similar adding the SSIM loss (exp #6/7 vs #3/4/5) significantly improves performance. The best performance is obtained by using all reconstruction loss components, without adversarial training (see 7, albeit adding a Vanilla GAN or LS-GAN performs similar, or at least within the initialisation variance).

Using a full component loss and an optimal generator backbone, we compare once more against other methods on the KITTI dataset (Tab. 3).

1.2. Cityscapes

Qualitative results are shown in Fig. 1 when trained on the cityscape dataset.

Table 1. Performance of models using different GAN variants. Cropping from Garg et al. (2016) was used for evaluation. For all results post-processing of disparity maps was performed.

	Loss Components					BN	GAN	ARD	SRD	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
	L1	LR	Disp	Occl	SSIM			lower is better				higher is better		
1						✓	V	0.810	12.442	18.245	1.999	0.002	0.008	0.020
1						✓	LS	0.893	13.826	18.816	2.468	0.000	0.000	0.000
1							W	0.813	12.310	18.119	1.932	0.001	0.003	0.011
2	✓							0.215	3.685	7.302	0.307	0.746	0.894	0.949
2	✓					✓		0.200	3.149	6.795	0.289	0.760	0.904	0.956
2	✓					✓	V	0.205	3.781	7.045	0.288	0.771	0.911	0.958
2	✓					✓	LS	0.190	2.826	6.612	0.281	0.766	0.909	0.959
2	✓						W	0.177	2.398	6.504	0.275	0.770	0.905	0.957
3	✓	✓						0.191	2.661	6.710	0.285	0.760	0.904	0.956
3	✓	✓				✓		0.162	1.755	5.954	0.253	0.789	0.922	0.966
3	✓	✓				✓	V	0.168	2.090	6.104	0.261	0.784	0.919	0.964
3	✓	✓				✓	LS	0.160	1.761	5.966	0.253	0.792	0.923	0.966
3	✓	✓					W	0.170	1.521	6.121	0.258	0.769	0.909	0.960
4	✓	✓	✓					0.200	3.155	7.039	0.295	0.758	0.900	0.953
4	✓	✓	✓			✓		0.163	1.842	5.978	0.253	0.791	0.922	0.966
4	✓	✓	✓			✓	V	0.165	1.907	6.094	0.258	0.787	0.920	0.964
4	✓	✓	✓			✓	LS	0.158	1.632	5.784	0.248	0.799	0.925	0.966
4	✓	✓	✓				W	0.160	1.427	6.179	0.259	0.772	0.908	0.959
5	✓	✓	✓	✓				0.204	3.399	6.983	0.295	0.760	0.901	0.953
5	✓	✓	✓	✓		✓		0.165	1.955	6.028	0.256	0.790	0.922	0.966
5	✓	✓	✓	✓		✓	V	0.196	3.182	6.582	0.282	0.778	0.911	0.957
5	✓	✓	✓	✓		✓	LS	0.174	2.236	6.137	0.263	0.785	0.917	0.962
5	✓	✓	✓	✓			W	0.161	1.557	6.191	0.260	0.776	0.910	0.960
6	✓	✓	✓		✓			0.142	1.200	5.694	0.239	0.809	0.927	0.967
6	✓	✓	✓		✓	✓	-	**0.132**	1.049	5.376	**0.224**	**0.822**	0.937	**0.974**
6	✓	✓	✓		✓	✓	V	0.135	1.052	5.428	0.229	0.818	0.935	0.972
6	✓	✓	✓		✓	✓	LS	0.135	1.051	5.417	0.227	0.819	0.936	0.972
6	✓	✓	✓		✓		W	0.152	1.357	6.003	0.249	0.788	0.917	0.963
7	✓	✓	✓	✓	✓			0.142	1.205	5.726	0.240	0.806	0.927	0.967
7	✓	✓	✓	✓	✓	✓		**0.132**	**1.035**	**5.370**	0.225	**0.822**	0.937	0.973
7	✓	✓	✓	✓	✓	✓	V	0.133	1.055	5.390	0.225	**0.822**	**0.938**	0.973
7	✓	✓	✓	✓	✓	✓	LS	0.134	1.090	5.447	0.226	0.820	0.937	0.973
7	✓	✓	✓	✓	✓		W	0.157	1.368	6.065	0.253	0.779	0.915	0.963
8	**Training set mean**							0.361	4.826	8.102	0.377	0.638	0.804	0.894

Table 2. Ablation study of loss components. All parameters γ weigh loss components from equation 1. α_{SSIM} is the ratio between L1 and SSIM.

Loss Weights γ					ARD	SRD	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
γ_{l1}	γ_s	γ_{lr}	γ_{disp}	γ_{occl}	lower is better				higher is better		
1.0	0.0	0.0	0.0	0.0	0.212	3.628	7.271	0.308	0.747	0.893	0.949
0.5	0.5	0.0	0.0	0.0	0.143	1.199	5.709	0.239	0.808	0.927	**0.968**
0.15	0.85	0.0	0.0	0.0	**0.142**	**1.186**	5.689	0.238	0.808	0.927	**0.968**
0.0	1.0	0.0	0.0	0.0	**0.142**	1.197	**5.637**	**0.237**	**0.811**	**0.928**	**0.968**
0.0	0.0	1.0	0.0	0.0	1.000	16.087	19.776	9.500	0.0	0.0	0.0
1.0	0.0	1.0	0.0	0.0	0.200	3.187	6.898	0.292	0.759	0.902	0.954
0.15	0.85	1.0	0.0	0.0	0.144	1.215	5.688	0.240	0.807	0.926	0.967
0.15	0.85	1.0	0.1	0.0	**0.142**	1.200	5.694	0.239	0.809	0.927	0.967
0.15	0.85	1.0	0.0	0.01	0.143	1.209	5.730	0.241	0.805	0.927	0.967
0.15	0.85	1.0	0.1	0.01	**0.142**	1.202	5.706	0.240	0.807	0.927	0.967
0.15	0.85	1.0	0.25	0.25	0.190	4.000	6.905	0.278	0.798	0.918	0.960
0.15	0.85	1.0	1.0	1.0	0.625	32.436	14.880	0.517	0.733	0.847	0.892

Table 3. Comparison of RN50 architecture with method of Godard et al. (2017), when pre-trained on CityScape (SC) dataset.

Method		Trained on	ARD	SRD	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
			lower is better				higher is better		
Supervised using Left-Right Correspondence									
Pilzer et al. (2018)		K	0.152	1.388	6.016	0.247	0.789	0.918	0.965
Godard et al. (2017)	VGG	K	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Our work	baseline	K	0.142	1.200	5.694	0.239	0.809	0.927	0.967
Our work	BN + S2	K	**0.128**	**1.026**	**5.313**	**0.222**	**0.830**	**0.939**	**0.973**
Godard et al. (2017)	RN50	CS + K	0.114	0.898	4.935	0.206	0.861	0.949	0.976
Our work	RN50+BN+S2	CS + K	**0.112**	**0.820**	**4.738**	**0.202**	**0.866**	**0.952**	**0.978**

References

Garg, R., BG, V.K., Carneiro, G., Reid, I., 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue, in: ECCV.

Godard, C., Mac Aodha, O., Brostow, G.J., 2017. Unsupervised monocular depth estimation with left-right consistency, in: CVPR.

Pilzer, A., Xu, D., Puscas, M., Ricci, E., Sebe, N., 2018. Unsupervised adversarial depth estimation using cycled generative networks, in: Int. Conf. on 3D Vision.

Yang, N., Wang, R., Stückler, J., Cremers, D., 2018. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry, in: ECCV.

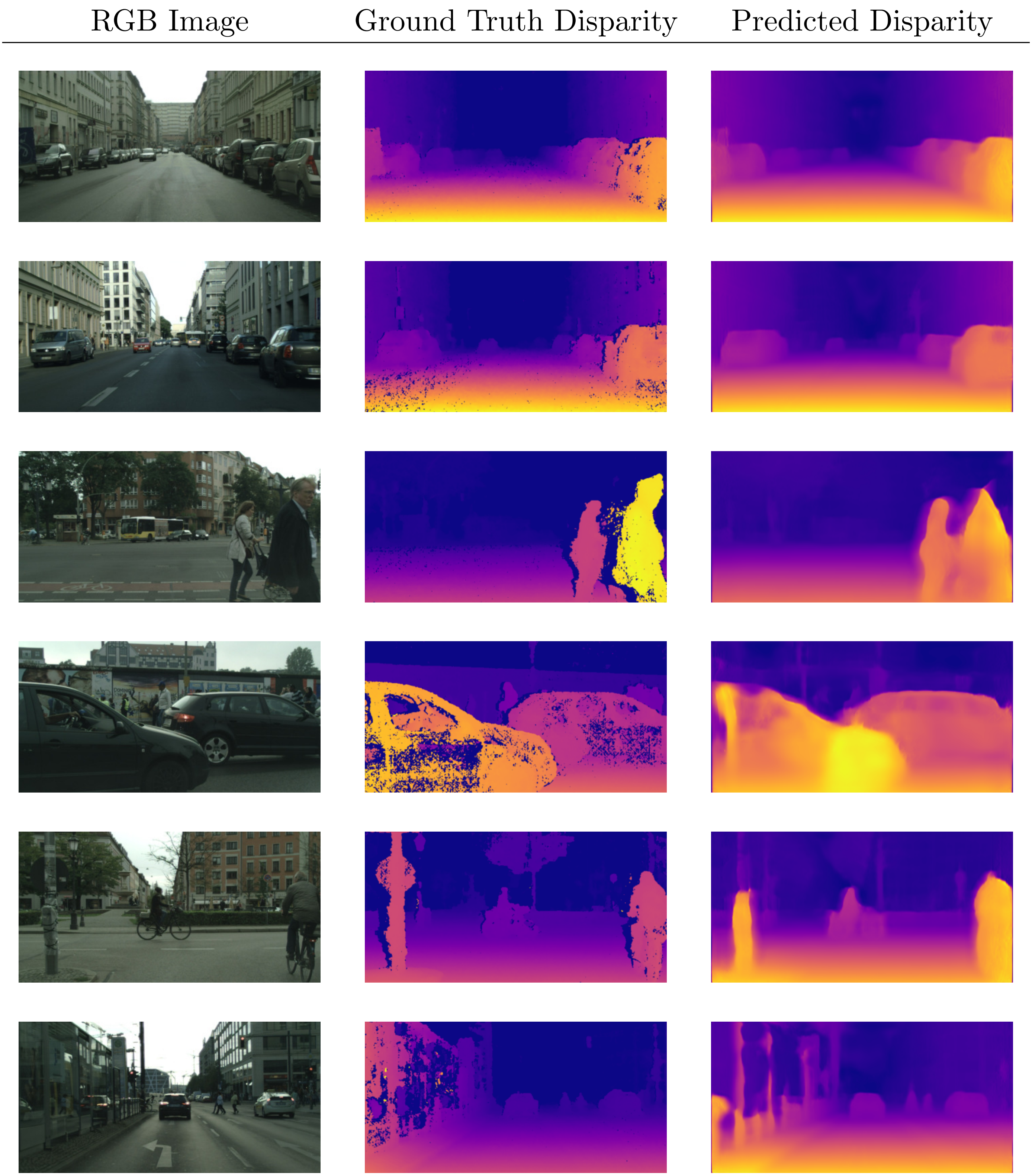

Fig. 1. Qualitative results on the CityScapes test set.