

Supplementary Material

Unsupervised Monocular Depth Estimation with Multi-Baseline Stereo

Quantitative analysis on Small Objects Dataset

For quantitative analysis, we prepare the ground truth for six images as shown in Figure 1. Following [1], we made the ground truth using COLMAP [6]. 8 images of each scene were taken with different viewpoints out of which one was used as the reference image. Similar to [4], we develop a 3D model of the scene to get depth map corresponding to the reference image, which is used as the ground truth. We compute both geometric and photometric depths. The pixels where these depths differ are not considered for evaluation. We convert obtained ground truth depth maps to inverse depth maps for evaluation. Since the scale of the ground truth depth maps obtained by COLMAP is undefined and it is not possible to recover absolute depth, we multiply the predicted disparity maps by scale factor s for evaluation similar to [3]. The scale factor s is defined as

$$s = \frac{\text{median}(d^*)}{\text{median}(d)}, \quad (1)$$

where d^* is the ground truth inverse depth and d is the predicted disparity. This scale factor is found for every predicted disparity map.

Method	Baseline	Lower the better				Higher the better		
		<i>AbsRel</i>	<i>SqRel</i>	<i>RMSE</i>	<i>RMSELog</i>	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth [2]+pp	2mm	0.1657	0.7817	5.786	0.371	0.843	0.901	0.910
Monodepth[2]+pp	10mm	0.1477	0.7142	5.752	0.372	0.869	0.898	0.910
3Net [5]+pp	2mm	0.1603	0.7241	5.780	0.367	0.851	0.900	0.910
3Net [5]+pp	10mm	0.1498	0.7142	5.806	0.392	0.862	0.894	0.907
monoResMatch [7]+pp	2mm	0.2100	0.9097	5.840	0.388	0.742	0.859	0.901
monoResMatch [7]+pp	10mm	0.1711	0.8603	5.783	0.400	0.856	0.889	0.904
Monodepth2 [3]	2mm	0.1728	0.8855	5.864	0.390	0.833	0.898	0.909
Monodepth2 [3]	10mm	0.1741	0.9999	5.876	0.400	0.863	0.893	0.902
Ours	2mm,10mm	0.1340	0.7005	5.708	0.366	0.872	0.901	0.910

Table 1: Evaluation on small objects dataset. For comparison, we train previous methods separately with 2 mm and 10 mm baseline stereo images. pp stands for post-processing.

The quantitative results are shown in Table 1. We report the results only with post-processing. The results show that our method outperforms other methods on all the metrics. The multi-baseline approach decreases the absolute relative error by 10% compared to the Monodepth trained on 10 mm

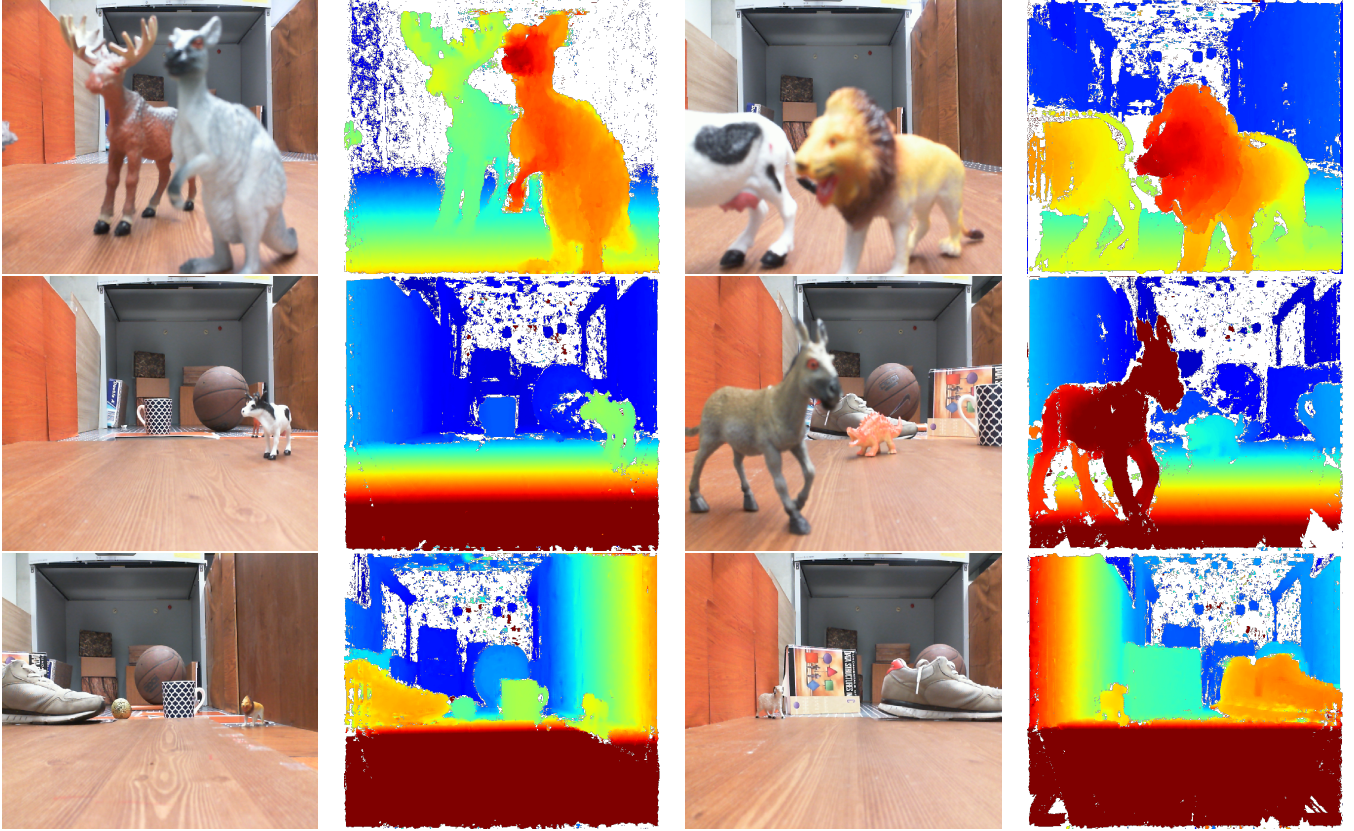


Figure 1: Ground truth images obtained using COLMAP [6]. White regions correspond to the pixels where geometric and photometric depth differs.

baseline, which is the second best performing method. The monoResMatch trained on 2 *mm* baseline performs worst among all the methods. Note that our and Monodepth2 results are reported without post-processing.

Choice of the weighting factor β

We set β based on the experiments on a small subset of the CARLA dataset. Depth prediction results at different values of β are shown in Table 2. Best results are obtained at $\beta = 0.85$.

β	SqRel
0.40	0.64
0.55	0.68
0.70	0.70
0.85	0.53
1.00	0.58

Table 2: Square relative errors at different values of β .

Network Architecture

Our network architecture is based on the encoder-decoder architecture of [2]. Unlike [2], we use three decoders for training. All the decoders use skip connections from the encoder. As per our experiments, using single decoder for training the network yields slightly worse results. Exponential linear unit (elu) is used as activation function for all the layers except disparity prediction layers. We use sigmoid non-linearity for predicting disparities. For up-convolution layers, we use nearest neighbour up-sampling followed by a convolution.

References

- [1] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T Barron. Learning single camera depth estimation using dual-pixels. In *IEEE International Conference on Computer Vision*, pages 7628–7637, 2019.
- [2] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.
- [3] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *IEEE International Conference on Computer Vision*, pages 3828–3838, 2019.
- [4] Saad Imran, Sikander Bin Mukarram, Muhammad Umar Karim Khan, and Chong-Min Kyung. Unsupervised deep learning for depth estimation with offset pixels. *Optics Express*, 28(6):8619–8639, 2020.
- [5] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *2018 International Conference on 3D Vision (3DV)*, pages 324–333. IEEE, 2018.
- [6] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [7] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019.