ENG: End-to-end Neural Geometry for Robust Depth and Pose Estimation using CNNs

An excerpt from the anonymous ECCV submission

Paper ID Anonymous

Abstract. Recovering structure and motion parameters given a image pair or a sequence of images is a well studied problem in computer vision. This is often achieved by employing Structure from Motion (SfM) or Simultaneous Localization and Mapping (SLAM) algorithms based on the real-time requirements. Recently, with the advent of Convolutional Neural Networks (CNNs) researchers have explored the possibility of using machine learning techniques to reconstruct the 3D structure of a scene and jointly predict the camera pose. In this work, we present a framework that achieves state-of-the-art performance on single image depth prediction for both indoor and outdoor scenes.

1 Dataset Evaluation Analysis

In this section we evaluate and analyse the relative performance on each dataset as well as correlations in the dataset and how they relate to overall performance.

1.1 NYUv2[1]

The dataset NYUv2[1] has been a popular benchmark for indoor depth estimation and semantic segmentation since the work of Eigen *et al.*[2]. We provide several qualitative and quantitative results from the evaluation of our approach in Figures 2, 3 and 4. This shows our strongest, median and worst performing images, as well as each predictions RMSE error in meters. This reveals two insights about our system's performance, and that is we perform stronger on images with closer median depths and that our largest errors occur when we incorrectly estimate the overall scale of the scene. The relationship to median depth is evident in Figure 1, where the RMSE is strongly correlated to the median scene depth. We also observe a similarly strong correlation in the performance of all three approaches, although our approach is overwhelmingly out performing the competitors.

What conclusions can we draw from these results? Well this is a rather clear result of the choice of error metric in ranking the results. In this case as we rank by the RMSE, we would expect higher depths to be the images with the largest error, as only either very large predictions or very large ground truth values can generate large RMSE values. This also indicates that our network tends to behave conservatively, estimating the scene is closer on average rather than further. This is probably a direct result of the depth value distribution in the training set, potentially biasing the depths towards the lower end.

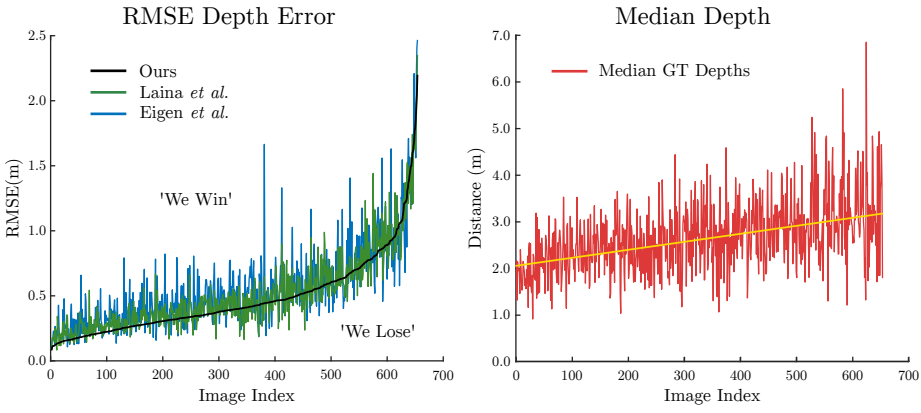

Fig. 1: *Left:* The RMSE error on each image of the test set, sorted by our performance on the NYUv2 dataset. We include two competing approaches, as well as marking which side of the line indicates we are better (*'We Win'*) and which we are worse (*'We Lose'*). *Right:* The median ground-truth depth of each image in the test set also sorted by our RMSE performance. We include an approximate trend-line to show the relationship between depth and RMSE in our system

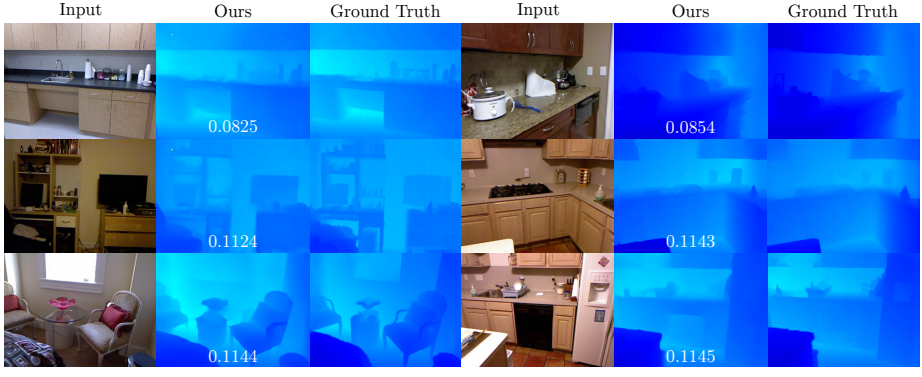

Fig. 2: The 6 highest performing images from the NYUv2 [1] testset, based on RMSE error. All images are of varying scenes, but contain lower median depth values on average.

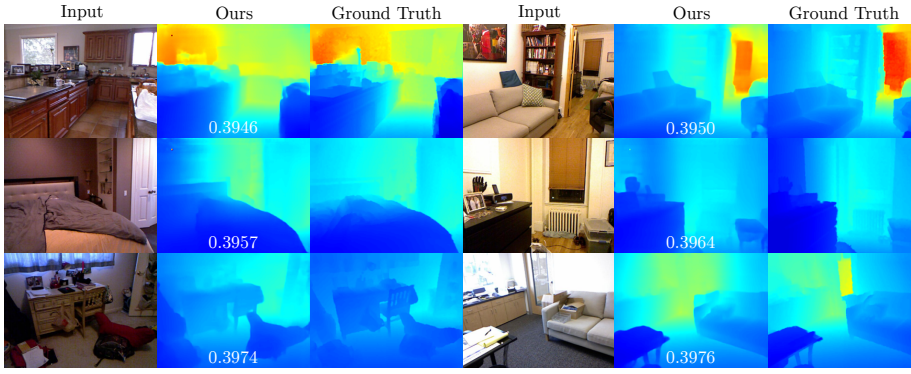

Fig. 3: The middle 6 images from the NYUv2 [1] testset, based on RMSE error. All images have RMSE values individually lower than the full testset (0.480m), indicating a small number of outliers, which is apparent in Figure 1

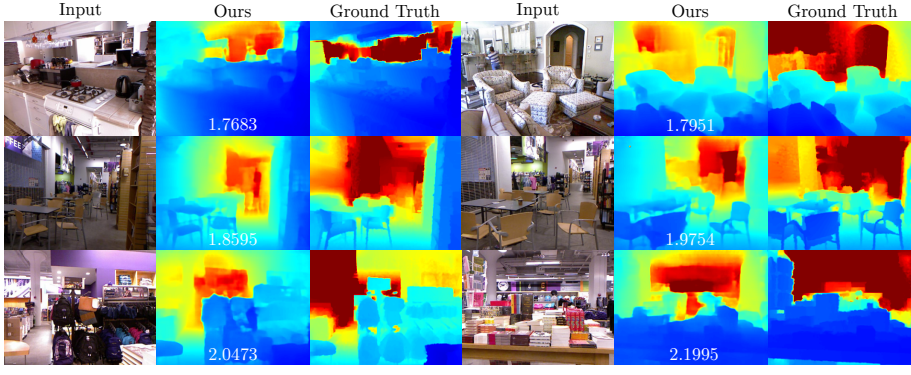

Fig. 4: The 6 lowest performing images from the NYUv2 [1] testset, based on RMSE error. In general these images contain higher median depth values, but the way in which our network gets it wrong appears to be in estimating the overall scene scale. This quantity is challenging to estimate, and we can observe qualitatively the system still produces believable relative depth estimates

1.2 KITTI [3]

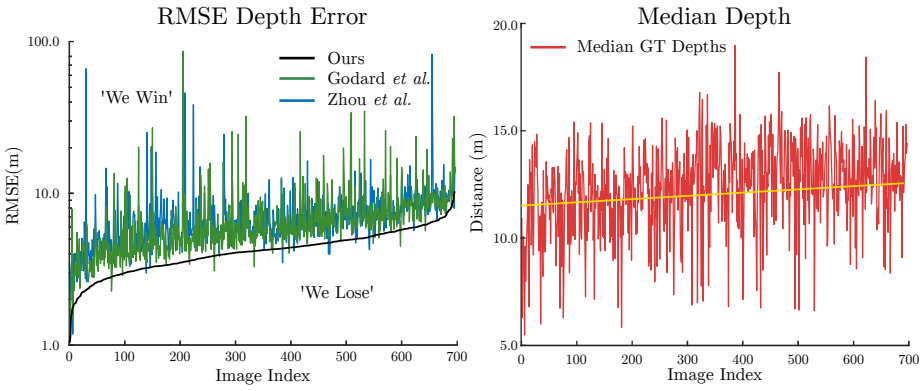

Fig. 5: *Left:* The RMSE error on each image of the test set, sorted by our performance on the KITTI dataset. We include two competing approaches, as well as marking which side of the line indicates we are better(*'We Win'*) and which we are worse (*'We Lose'*). *Right:* The median ground-truth depth of each image in the test set also sorted by our RMSE performance. We include an approximate trend-line to show the relationship between depth and RMSE in our system

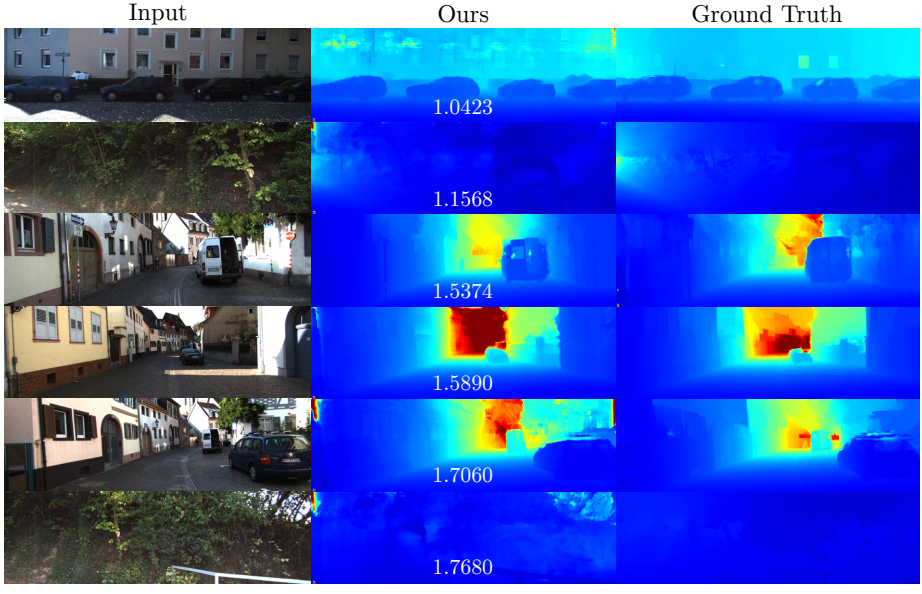

Fig. 6: The highest performing 6 images from the KITTI [3] testset, based on RMSE error. Surprisingly not all of these contain a great deal of scale context, in particular rows 2 and 6, where they face a dirt ramp, which is atypical of the predominantly road facing dataset. This indicates strongly that the approach is genuinely learning about the geometry of the scenes

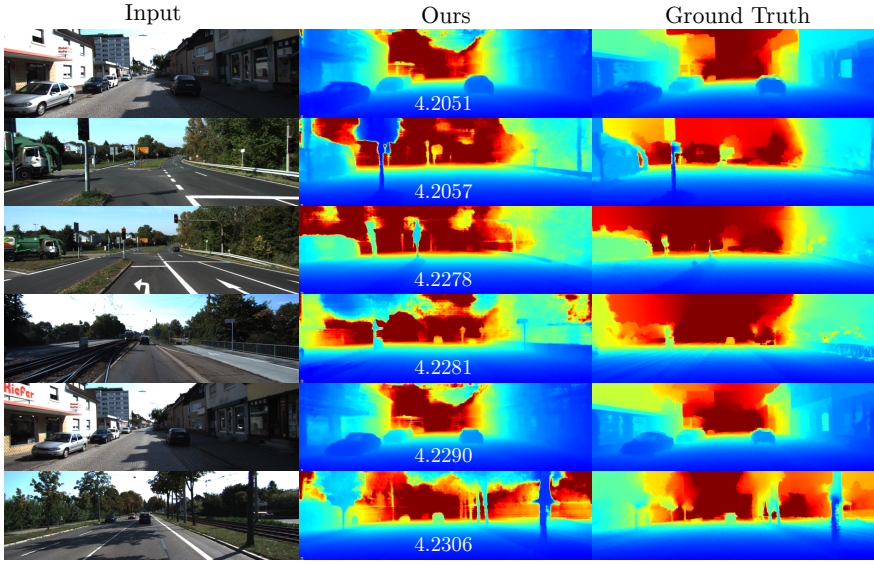

Fig. 7: The middle 6 performing images from the KITTI [3] testset, based on RMSE error. The RMSE values are hovering around the value achieved for the dataset and represent the typical performance. Note the systems ability to estimate depth in the top half of the scene, which never receives a ground truth training signal, as the LIDAR only scans below the horizon line

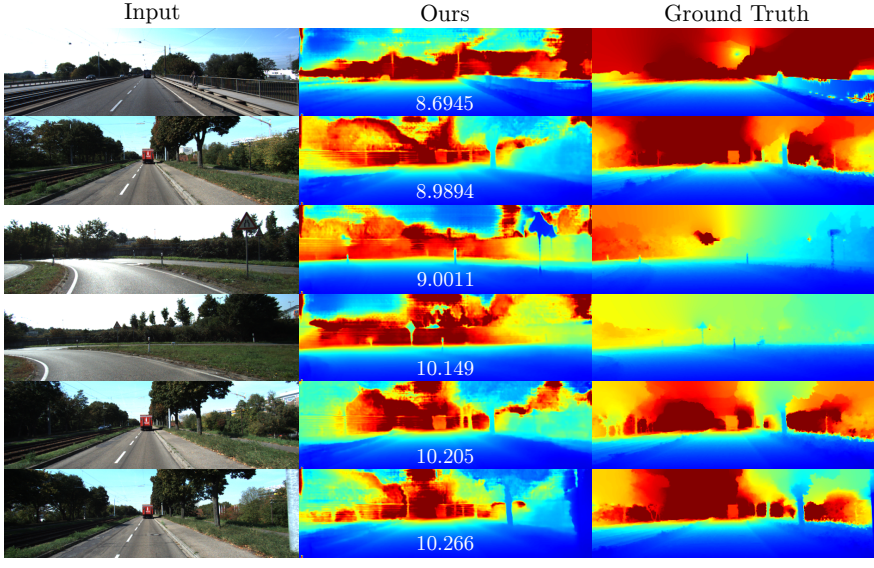

Fig. 8: The lowest 6 performing images from the KITTI [3] testset, based on RMSE error. Although the RMSE of each of these images is comparitively high, the depth predictions produced are convincing qualitatively

Our most impressive performance is perhaps on the KITTI benchmark dataset [3]. Where as shown in Figure 5 (*left*), we consistently out-perform the competing approaches on almost all test images. The scale of the depth error had to be changed to $\log_{10}$ in order to capture the full range of errors. This could be because the competing approaches estimate inverse depth/disparity and invert the predicted values to compute their loss function. This can lead to unstable performance on large distances, due to the non-linearity of this section, as opposed to our approach which is linear to all depths.

Again we include the analysis of the median depths sorted against the RMSE error in Figure 5 (*right*), as we did for NYUv2. In this case the relationship between error and depth is largely reduced, this is most likely due to the nature of the dataset which contains a very similar spread of data for most images in the training set, as they film similar scenarios. However the relationship is still visible in Figure 6, where these scenes contain comparitively low depth values, indicating again our system behaves conservatively in estimating depths.

1.3 Quantitative Analysis

We summarise the results of evaluating our single-image depth estimation of the datasets NYUv2[1], RGB-D[4] and KITTI[3] in Tables 1, 2 and 3 respectively. We quantitatively evaluate against previous state-of-the-art approaches using the standard metrics proposed in [5].

We show significant improvement across all datasets using our baseline approach (*Ours(baseline)*), validating the choice of architecture used in this work. Further to this, we demonstrate a consistent improvement across all datasets when we infer using the fully end-to-end trained network (*Ours(full)*). Most notably in Tables 2 and 3 for which ground truth pose data was available for training. This validates our approach for improving single image depth estimation performance, and demonstrates a network can be improved by enforcing more geometric priors on the loss functions.

Table 1: The performance of several approaches evaluated on single-image depth estimation using the standard testset of NYUv2[1] proposed in [2]

Method	*lower better*			*higher better*		
	RMS_{lin}	RMS_{ln}	Rel_{abs}	δ	δ^2	δ^3
Eigen_{vgg} [2]	0.641	0.214	0.16	76.9%	95.0%	98.8%
Laina *et al.*[6]	0.573	0.195	0.13	81.1%	95.3%	98.8%
Kendall *et al.* [7]	0.506	-	**0.110**	81.7%	95.9%	98.9%
Ours (baseline)	0.487	0.164	0.113	86.7%	97.7%	99.4%
Ours (full)	**0.478**	**0.161**	0.111	**87.2%**	**97.8%**	**99.5%**

Additionally we include qualitative results for NYUv2[1] and KITTI[3] in Figure 9 and 10 respectively. Each of which illustrates a noticeable improvement over previous methods. We also demonstrate that the improvement is beyond the numbers, as our approach generates more convincing depths even when the RMSE may be higher, as is the

Table 2: The performance of previous state-of-the-art approaches evaluated on the standard testset of the KITTI dataset [3]

Cap	Method	*lower better*			*higher better*		
		RMS_{lin}	RMS_{ln}	Rel_{abs}	δ	δ^2	δ^3
0-80m	SFM-learner[8]	6.856	0.283	0.208	67.8%	88.5%	95.7%
	Godard *et al.*[9]	4.935	0.206	0.141	86.1%	94.9%	97.6%
	Kuznietsov *et al.* [10]	4.621	0.189	0.113	86.2%	96.0%	98.6%
	Ours (baseline)	4.394	0.178	0.095	89.4%	96.6%	98.6%
	Ours (full)	**4.301**	**0.173**	**0.096**	**89.5%**	**96.8%**	**98.7%**
0-50m	SFM-learner[8]	5.181	0.264	0.201	69.6%	90.0%	96.6%
	Garg *et al.* [11]	5.104	0.273	0.169	74.0%	90.4%	96.2%
	Godard *et al.* [9]	3.729	0.194	0.108	87.3%	95.4%	97.9%
	Kuznietsov *et al.*[10]	3.518	0.179	0.108	87.5%	96.4%	98.8%
	Ours(baseline)	3.359	0.168	0.092	90.5%	97.0%	98.8%
	Ours(full)	**3.284**	**0.164**	**0.092**	**90.6%**	**97.1%**	**98.9%**

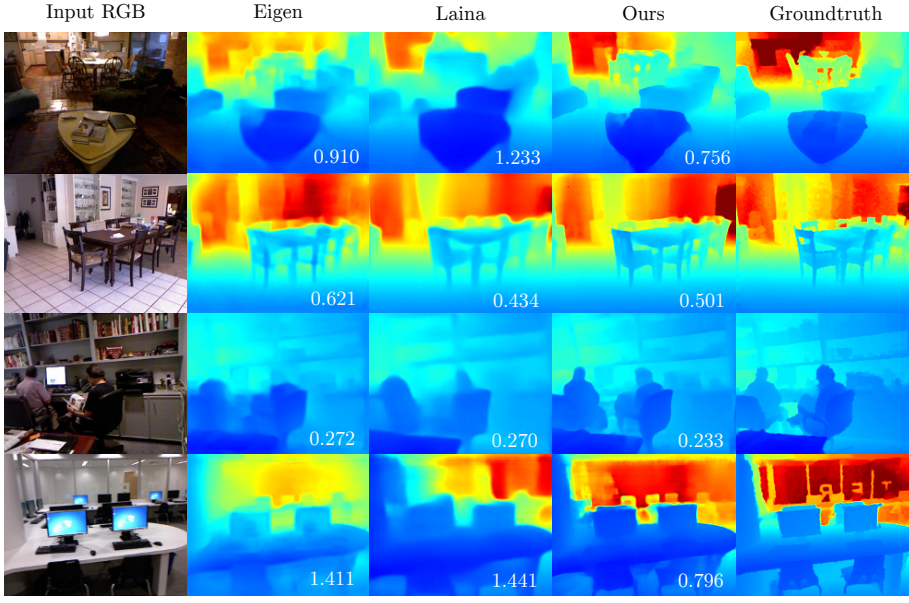

Fig. 9: Resulting single image depth estimation for several approaches and ours against the ground truth on the dataset NYUv2[1]. The RMSE for each prediction is included

case in the second row of Figure 9, where [6] computes a lower RMSE. More impressive still are the results in Figure 10, where we compare against previous approaches that are both trained on much larger training sets than our own and still show noticeable qualitative and quantitative improvements.

Table 3: The performance of previous state-of-the-art approaches on a randomly selected subset of the frames from the RGB-D dataset [4]. We post separate entries for DeMoN(est) and DeMoN(gt), former is scaled by the estimated scale of their system while the latter is scaled by the median groundtruth depth

Method	*lower better*				*higher better*		
	RMS_{lin}	RMS_{log}	Rel_{abs}	Rel_{sqr}	δ	δ^2	δ^3
Laina *et al.*[6]	1.275	0.481	0.189	0.371	75.3%	89.1%	91.8%
DeMoN(est)[12]	2.980	0.910	1.413	5.109	21.0%	36.6%	48.9%
DeMoN(gt)[12]	1.584	0.555	0.301	0.581	52.7%	70.7%	80.7%
Ours(baseline)	1.068	0.353	0.128	0.236	86.9%	92.2%	93.5%
Ours(full)	**0.996**	**0.329**	**0.108**	**0.194**	**90.3%**	**93.6%**	**94.5%**

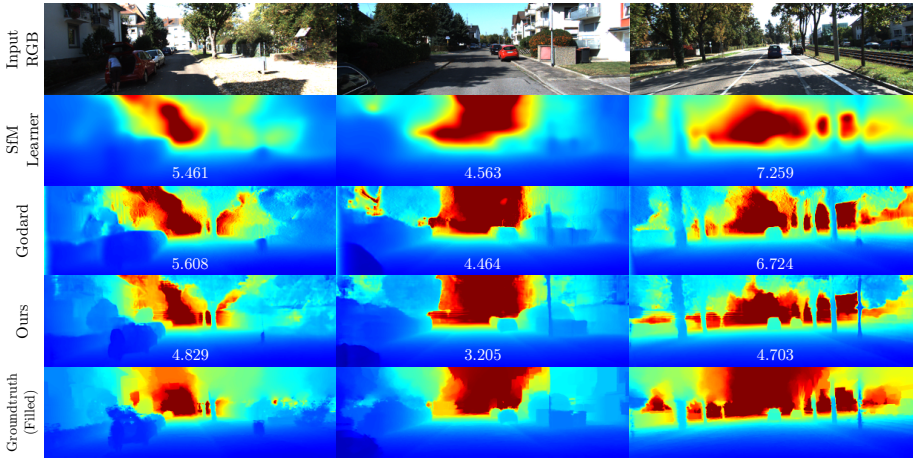

Fig. 10: The resulting single image depth estimation for several approaches including SfM-Learner[8], Godard[9] and Ours against a ground truth filled using [13] on the testset of the KITTI dataset [3]. We include the RMSE values for each methods prediction. Filled depths are included for visualisation purposes during evaluation the predictions are evaluated against the sparse velodyne ground truth data.

References

1. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European Conference on Computer Vision (ECCV). (2012) 1–14
2. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: IEEE International Conference on Computer Vision (ICCV). (2015) 2650–2658
3. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. International Journal of Robotics Research (IJRR) (2013)
4. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: International Conference on Intelligent Robot Systems (IROS). (2012)
5. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems (NIPS). (2014) 2366–2374
6. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: International Conference on 3D Vision (3DV). (2016) 239–248
7. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems (NIPS). (2017) 5580–5590
8. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
9. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
10. Kuznietsov, Y., Stückler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
11. Garg, R., BG, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European Conference on Computer Vision (ECCV). (2016) 740–756
12. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
13. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. In: ACM Transactions on Graphics. Volume 23. (2004) 689–694