Disentangling Object Motion and Occlusion for Unsupervised Multi-frame Monocular Depth Supplementary Materials

Ziyue Feng[1], Liang Yang[2], Longlong Jing[2], Haiyan Wang[2], YingLi Tian[2], and Bing Li[1]

[1] Clemson University,
[2] City University of New York

This document contains the supplementary materials for "Disentangling Object Motion and Occlusion for Unsupervised Multi-frame Monocular Depth". Code is available at `https://github.com/AutoAILab/DynamicDepth`

1 Additional Implementation Details

Occlusion-aware Re-projection Loss: We obtain the exact occlusion mask O and visible mask V from our DOMD module M_o, our Occlusion-aware Re-projection Loss L_{or} always choose the non-occluded source frame pixels for photo-metric error.

$$L_{or} = \frac{1}{|I_t - (O_{t-1} \cap O_{t+1})|} \sum_{i \in I_t} E^i_{or}, \tag{1}$$

$$E^i_{or} = \begin{cases} EO^i_{t-1}, & \text{if } I^i_{t-1} \in V_{t-1}, I^i_{t+1} \in O_{t+1}, \\ EO^i_{t+1}, & \text{if } I^i_{t-1} \in O_{t-1}, I^i_{t+1} \in V_{t+1}, \\ \min(EO^i_{t-1}, EO^i_{t+1}), & \text{if } I^i_{t-1} \in V_{t-1}, I^i_{t+1} \in V_{t+1}, \\ 0, & \text{if } I^i_{t-1} \in O_{t-1}, I^i_{t+1} \in O_{t+1}. \end{cases} \tag{2}$$

Depth Prior Net: Our Depth Prior Net θ_{DPN} consists of a depth encoder and a depth decoder. We use an ImageNet [4] pre-trained ResNet18 [13] as backbone for depth encoder, which has 4 pyramidal scales. Features in each scale are fed to the depth decoder by several UNet [31] style skip connections. The depth decoder consists of multiple convolution layers for the encoder feature fusion and nearest interpolations for up-sampling.

Pose Net: Our Pose Net shares a similar architecture as our Depth Prior Net, but it outputs a 6-degree-of-freedom camera ego-motion vector P_o instead of the depth map.

DOMD: Our Dynamic Object Motion Disentanglement (DOMD) module projects the object image patches C_t to C^d_{t-1} to replace C_{t-1} to disentangle the object motion. The projection is based on the depth prior prediction D^{pr}_t, known camera intrinsics K, and camera ego-motion prediction P_o. We do not need instance-level masks and inter-frame correspondences, all dynamic objects are projected together at once. We use an off-the-shelf semantic segmentation model EffcientPS [25] to provide the dynamic category segmentation masks. We define the dynamic category as follows: {person, rider, car, truck, bus, caravan, trailer, motorcycle, bicycle}.

Cost Volume: We pre-define 96 different depth hypothesis bins and reduce the channel number to 1. The cost volume is constructed at the third scale which is in 48×160 resolution, resulting in an $CV \in R^{96\times160\times48\times1}$. Our cost volume only consumes $2.8MB$ memory when using Float32 data type.

Depth Encoder and Decoder: Our depth encoder and decoder in the multi-frame model θ_{MF} shares the same architecture with the Depth Prior Net θ_{DPN}. The Occlusion-aware Cost Volume is integrated at the third scale of the encoder.

Training: We use frames $\{I_{t-1}, I_t, I_{t+1}\}$ for training and $\{I_{t-1}, I_t\}$ for testing. Our model is trained using an Adam [15] optimizer with a learning rate of 10^{-4} for 10 epochs, which takes about 10 hours on a single Nvidia A100 GPU.

Evaluation Metrics: Following the state-of-the-art methods [8,10,32], we use Absolute Relative Error (Abs Rel), Squared Relative Error (Sq Rel), Root Mean Squared Error (RMSE), Root Mean Squared Log Error (RMSE$_{log}$), and δ_1, δ_2, δ_3 as the metrics to evaluate the depth prediction performance. These metrics are formulated as:

$$\text{AbsRel} = \frac{1}{n}\sum_i \frac{|p_i - g_i|}{g_i}, \qquad \text{SqRel} = \frac{1}{n}\sum_i \frac{(p_i - g_i)^2}{g_i},$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_i (p_i - g_i)^2}, \qquad \text{RMSE}_{\text{log}} = \sqrt{\frac{1}{n}\sum_i (\log p_i - \log g_i)^2},$$

$\delta_1, \delta_2, \delta_3 = \%\ of\ thresh < 1.25, 1.25^2, 1.25^3$, where g and p are the depth values of ground truth and prediction in meters, $thresh = \max(\frac{g}{p}, \frac{p}{g})$.

2 Additional Quantitative Results

2.1 KITTI Benchmark Scores

The original Eigen [5] split of KITTI [24] dataset uses the re-projected single-frame raw LIDAR points as ground truth for evaluation, which may contain outliers such as reflection on transparent objects. We only reported results with this original ground truth in the main paper since it is the most widely used. Jonas *et al.* [34] introduced a set of high-quality ground truth depth maps for the KITTI dataset, accumulates 5 consecutive frames to form the denser ground truth depth map, and removed the outliers. This improved ground truth depth is provided for 652 (or 93%) of the 697 test frames contained in the Eigen test split [5]. We evaluate our method on these 652 improved ground truth frames and compare with existing state-of-art published methods in Table 1. Following the convention, we clip the predicted depths to 80 meters to match the Eigen evaluation. Methods are ranked by the Absolute Relative Error. Our method outperforms all existing state-of-the-art methods, even some stereo-based and supervised methods.

2.2 Full Quantitative Results

Due to the space limitation, we only show a part of the quantitative comparison of depth prediction in the main paper. Here we show an extensive comparison to existing state-of-the-art methods on the KITTI [24] and Cityscapes [3] dataset in Table. 2. Following the convention, methods are sorted by the Abs Rel, which is the relative error with the ground truth. Our method outperforms all other state-of-the-art methods by a large margin, especially on the challenging Cityscapes [3] dataset, which contains significantly more dynamic objects. Our method even outperformed some stereo-based and supervised methods on the KITTI dataset. Note that all KITTI results in this section are based on the widely-used original [24] ground truth, which generates much greater error than the improved [34] ground truth.

3 Additional Qualitative Results

Fig 1 shows a full version of the qualitative results and Fig 2 shows an additional set of comparisons. We compare our results with other state-of-the-art methods. The I^d_{t-1} image disentangled the dynamic object motion to solve the mismatch problem. As shown in the histograms, most pixels of our method have lower depth error. Our method has lighter red color in the error map which indicates lower depth errors. The dynamic object area depths are projected to 3D point clouds and compared with ground truth point clouds, our prediction matches the ground truth significantly better.

Method	Training	WxH	The lower the better				The higher the better		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhan FullNYU [43]	Sup	608 x 160	0.130	1.520	5.184	0.205	0.859	0.955	0.981
Kuznietsov et al. [16]	Sup	621 x 187	0.089	0.478	3.610	0.138	0.906	0.980	0.995
DORN [6]	Sup	513 x 385	0.072	0.307	2.727	0.120	0.932	0.984	0.995
Monodepth [7]	S	512 x 256	0.109	0.811	4.568	0.166	0.877	0.967	0.988
3net [29] (VGG)	S	512 x 256	0.119	0.920	4.824	0.182	0.856	0.957	0.985
3net [29] (ResNet 50)	S	512 x 256	0.102	0.675	4.293	0.159	0.881	0.969	0.991
SuperDepth [27]	S	1024 x 384	0.090	0.542	3.967	0.144	0.901	0.976	0.993
Monodepth2 [8]	S	640 x 192	0.085	0.537	3.868	0.139	0.912	0.979	0.993
EPC++ [21]	S	832 x 256	0.123	0.754	4.453	0.172	0.863	0.964	0.989
SfMLearner [46]	M	416 x 128	0.176	1.532	6.129	0.244	0.758	0.921	0.971
Vid2Depth [22]	M	416 x 128	0.134	0.983	5.501	0.203	0.827	0.944	0.981
GeoNet [42]	M	416 x 128	0.132	0.994	5.240	0.193	0.833	0.953	0.985
DDVO [35]	M	416 x 128	0.126	0.866	4.932	0.185	0.851	0.958	0.986
Ranjan [30]	M	832 x 256	0.123	0.881	4.834	0.181	0.860	0.959	0.985
EPC++ [21]	M	832 x 256	0.120	0.789	4.755	0.177	0.856	0.961	0.987
Johnston *et al.* [14]	M	640 x 192	0.081	0.484	3.716	0.126	0.927	0.985	0.996
Monodepth2 [8]	M	640 x 192	0.090	0.545	3.942	0.137	0.914	0.983	0.995
Packnet-SFM [10]	M	640 x 192	0.078	0.420	3.485	0.121	0.931	0.986	0.996
Patil *et al.*[26]	M	640 x 192	0.087	0.495	3.775	0.133	0.917	0.983	0.995
Wang *et al.*[37]	M	640 x 192	0.082	0.462	3.739	0.127	0.923	0.984	0.996
ManyDepth [39]	M	640 x 192	0.070	0.399	3.455	0.113	0.941	0.989	0.997
DynamicDepth	M	640 x 192	**0.068**	**0.362**	**3.454**	**0.111**	**0.943**	**0.991**	**0.998**

Table 1. KITTI Evaluation on Improved Ground Truth [34]: Following the convention, methods in each category are sorted by the Abs Rel, which is the relative error with the ground truth. Best methods are in **bold**. Our method out-performs all other state-of-the-art methods, even some stereo-based and supervised methods.

Legend: Sup – Supervised by ground truth depth S – Stereo M – Monocular

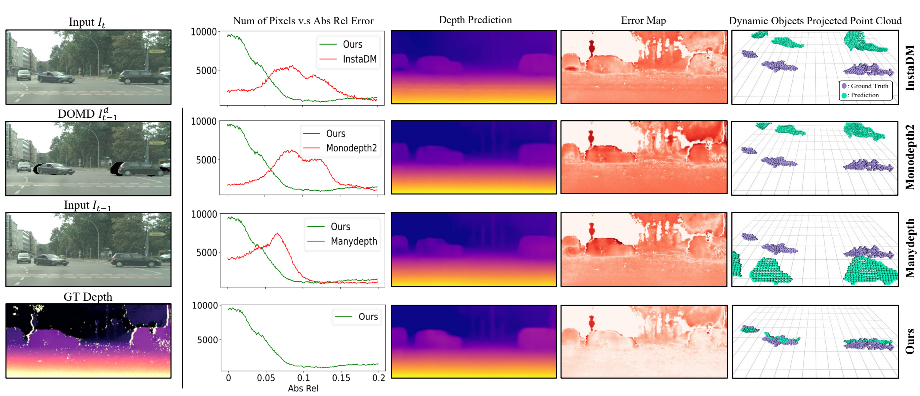

Fig. 1. Full Qualitative visualization: The left column shows the input image frames and our disentangled image I^d_{t-1}, later columns show the comparison with other state-of-the-art methods. In the histograms, most pixels of our method has lower depth error. In the error map, our method has lighter red color which indicates lower depth errors. We project the dynamic object area depths to 3D point clouds and compare them with ground truth point clouds in the last column. Our prediction matches the ground truth significantly better.

	Method	Training	WxH	The lower the better				The higher the better		
				Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
KITTI Original	Zhan FullNYU [43]	Sup	608 x 160	0.135	1.132	5.585	0.229	0.820	0.933	0.971
	Kuznietsov et al. [16]	Sup	621 x 187	0.113	0.741	4.621	0.189	0.862	0.960	0.986
	Gur et al. [12]	Sup	416 x 128	0.110	0.666	4.186	0.168	0.880	0.966	0.988
	Dorn [6]	Sup	513 x 385	0.099	0.593	3.714	0.161	0.897	0.966	0.986
	MonoDepth [7]	S	512 x 256	0.133	1.142	5.533	0.230	0.830	0.936	0.970
	MonoDispNet [40]	S	512 x 256	0.126	0.832	4.172	0.217	0.840	0.941	0.973
	MonoResMatch [33]	S	1280 x 384	0.111	0.867	4.714	0.199	0.864	0.954	0.979
	MonoDepth2 [8]	S	640 x 192	0.107	0.849	4.764	0.201	0.874	0.953	0.977
	UnDeepVO [20]	S	512 x 128	0.183	1.730	6.570	0.268	-	-	-
	DFR [44]	S	608 x 160	0.135	1.132	5.585	0.229	0.820	0.933	0.971
	EPC++ [21]	S	832 x 256	0.128	0.935	5.011	0.209	0.831	0.945	0.979
	DepthHint [38]	S	640 x 192	0.100	0.728	4.469	0.185	0.885	0.962	0.982
	FeatDepth [32]	S	640 x 192	0.099	0.697	4.427	0.184	0.889	0.963	0.982
	SfMLearner [46]	M	416 x 128	0.208	1.768	6.958	0.283	0.678	0.885	0.957
	Vid2Depth [22]	M	416 x 128	0.163	1.240	6.220	0.250	0.762	0.916	0.968
	LEGO [41]	M	416 x 128	0.162	1.352	6.276	0.252	0.783	0.921	0.969
	GeoNet [42]	M	416 x 128	0.155	1.296	5.857	0.233	0.793	0.931	0.973
	DDVO [36]	M	416 x 128	0.151	1.257	5.583	0.228	0.810	0.936	0.974
	DF-Net [47]	M	576 x 160	0.150	1.124	5.507	0.223	0.806	0.933	0.973
	Ranjan *et al.*[30]	M	832 x 256	0.148	1.149	5.464	0.226	0.815	0.935	0.973
	EPC++ [21]	M	832 x 256	0.141	1.029	5.350	0.216	0.816	0.941	0.976
	Struct2depth (M) [1]	M	416 x 128	0.141	1.026	5.291	0.215	0.816	0.945	0.979
	SIGNet [23]	M	416 x 128	0.133	0.905	5.181	0.208	0.825	0.947	0.981
	Li *et al.*[19]	M	416 x 128	0.130	0.950	5.138	0.209	0.843	0.948	0.978
	Videos in the wild [9]	M	416 x 128	0.128	0.959	5.230	0.212	0.845	0.947	0.976
	DualNet [45]	M	1248 x 384	0.121	0.837	4.945	0.197	0.853	0.955	0.982
	SuperDepth [27]	M	1024 x 384	0.116	1.055	-	0.209	0.853	0.948	0.977
	Monodepth2 [8]	M	640 x 192	0.115	0.903	4.863	0.193	0.877	0.959	0.981
	Lee *et al.* [18]	M	832 x 256	0.114	0.876	4.715	0.191	0.872	0.955	0.981
	InstaDM [17]	M	832 x 256	0.112	0.777	4.772	0.191	0.872	0.959	0.982
	Patil *et al.*[26]	M	640 x 192	0.111	0.821	4.650	0.187	0.883	0.961	0.982
	Packnet-SFM [10]	M	640 x 192	0.111	0.785	4.601	0.189	0.878	0.960	0.982
	Wang *et al.*[37]	M	640 x 192	0.106	0.799	4.662	0.187	0.889	0.961	0.982
	Johnston *et al.* [14]	M	640 x 192	0.106	0.861	4.699	0.185	0.889	0.962	0.982
	FeatDepth [32]	M	640 x 192	0.104	0.729	4.481	0.179	0.893	0.965	0.984
	Guizilini *et al.*[11]	M	640 x 192	0.102	0.698	4.381	0.178	0.896	0.964	0.984
	ManyDepth [39]	M	640 x 192	0.098	0.770	4.459	0.176	**0.900**	**0.965**	0.983
	DynamicDepth	M	640 x 192	**0.096**	**0.720**	**4.458**	**0.175**	0.897	0.964	**0.984**
Cityscapes	Pilzer *et al.*[28]	M	512 x 256	0.240	4.264	8.049	0.334	0.710	0.871	0.937
	Struct2Depth 2 [2]	M	416 x 128	0.145	1.737	7.280	0.205	0.813	0.942	0.976
	Monodepth2 [8]	M	416 x 128	0.129	1.569	6.876	0.187	0.849	0.957	0.983
	Videos in the Wild [9]	M	416 x 128	0.127	1.330	6.960	0.195	0.830	0.947	0.981
	Li *et al.*[19]	M	416 x 128	0.119	1.290	6.980	0.190	0.846	0.952	0.982
	Lee *et al.* [18]	M	832 x 256	0.116	1.213	6.695	0.186	0.852	0.951	0.982
	InstaDM [17]	M	832 x 256	0.111	1.158	6.437	0.182	0.868	0.961	0.983
	Struct2Depth 2 [2]	M	416 x 128	0.151	2.492	7.024	0.202	0.826	0.937	0.972
	ManyDepth [39]	M	416 x 128	0.114	1.193	6.223	0.170	0.875	0.967	0.989
	DynamicDepth	M	416 x 128	**0.103**	**1.000**	**5.867**	**0.157**	**0.895**	**0.974**	**0.991**

Table 2. Depth Prediction on KITTI and Cityscapes Dataset. Following the convention, methods in each category are sorted by the Abs Rel, which is the relative error with the ground truth. Best methods are in **bold**. Our method out-performs all other state-of-the-art methods by a large margin especially on the challenging Cityscapes [3] dataset, which contains significantly more dynamic objects. Our method even outperformed some stereo based and supervised methods on KITTI dataset. Note that all KITTI result in this table are based on the widely-used original [24] ground truth, which generates much greater error than the improved [34] ground truth.

Legend: Sup – Supervised by ground truth depth S – Stereo M – Monocular

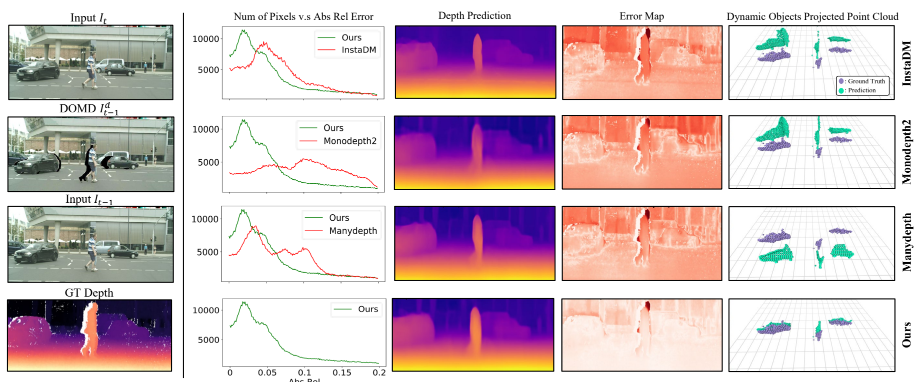

Fig. 2. Additional Qualitative visualization: The left column shows the input image frames and our disentangled image I^d_{t-1}, later columns show the comparison with other state-of-the-art methods. In the histograms, most pixels of our method has lower depth error. In the error map, our method has lighter red color which indicates lower depth errors. We project the dynamic object area depths to 3D point clouds and compare them with ground truth point clouds in the last column. Our prediction matches the ground truth significantly better.

References

1. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: AAAI (2019)
2. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Unsupervised monocular depth and ego-motion learning with structure and semantics. In: CVPR Workshops (2019)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
6. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2002–2011 (2018)
7. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
8. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3828–3838 (2019)
9. Gordon, A., Li, H., Jonschkowski, R., Angelova, A.: Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In: ICCV (2019)
10. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3D packing for self-supervised monocular depth estimation. In: CVPR (2020)
11. Guizilini, V., Hou, R., Li, J., Ambrus, R., Gaidon, A.: Semantically-guided representation learning for self-supervised monocular depth. In: ICLR (2020)
12. Gur, S., Wolf, L.: Single image depth estimation trained via depth from defocus cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7683–7692 (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Johnston, A., Carneiro, G.: Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In: CVPR (2020)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Kuznietsov, Y., Stuckler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6647–6655 (2017)
17. Lee, S., Im, S., Lin, S., Kweon, I.S.: Learning monocular depth in dynamic scenes via instance-aware projection consistency. In: 35th AAAI Conference on Artificial Intelligence/33rd Conference on Innovative Applications of Artificial Intelligence/11th Symposium on Educational Advances in Artificial Intelligence. pp. 1863–1872. ASSOC ADVANCEMENT ARTIFICIAL INTELLIGENCE (2021)

18. Lee, S., Rameau, F., Pan, F., Kweon, I.S.: Attentive and contrastive learning for joint depth and motion field estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4862–4871 (2021)
19. Li, H., Gordon, A., Zhao, H., Casser, V., Angelova, A.: Unsupervised monocular depth learning in dynamic scenes. In: CoRL (2020)
20. Li, R., Wang, S., Long, Z., Gu, D.: Undeepvo: Monocular visual odometry through unsupervised deep learning. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 7286–7291. IEEE (2018)
21. Luo, C., Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R., Yuille, A.: Every pixel counts++: Joint learning of geometry and motion with 3D holistic understanding. PAMI (2019)
22. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and egomotion from monocular video using 3D geometric constraints. In: CVPR (2018)
23. Meng, Y., Lu, Y., Raj, A., Sunarjo, S., Guo, R., Javidi, T., Bansal, G., Bharadia, D.: Signet: Semantic instance aided unsupervised 3d geometry perception. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 9810–9820. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.01004, http://openaccess.thecvf.com/content_CVPR_2019/html/Meng_SIGNet_Semantic_Instance_Aided_Unsupervised_3D_Geometry_Perception_CVPR_2019_paper.html
24. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
25. Mohan, R., Valada, A.: Efficientps: Efficient panoptic segmentation. International Journal of Computer Vision **129**(5), 1551–1579 (2021)
26. Patil, V., Van Gansbeke, W., Dai, D., Van Gool, L.: Don't forget the past: Recurrent depth estimation from monocular video. IEEE Robotics and Automation Letters **5**(4), 6813–6820 (2020)
27. Pillai, S., Ambrus, R., Gaidon, A.: Superdepth: Self-supervised, super-resolved monocular depth estimation. In: ICRA (2019)
28. Pilzer, A., Xu, D., Puscas, M.M., Ricci, E., Sebe, N.: Unsupervised adversarial depth estimation using cycled generative networks. In: 3DV (2018)
29. Poggi, M., Tosi, F., Mattoccia, S.: Learning monocular depth estimation with unsupervised trinocular assumptions. In: 3DV (2018)
30. Ranjan, A., Jampani, V., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: CVPR (2019)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
32. Shu, C., Yu, K., Duan, Z., Yang, K.: Feature-metric loss for self-supervised learning of depth and egomotion. In: European Conference on Computer Vision. pp. 572–588. Springer (2020)
33. Tosi, F., Aleotti, F., Poggi, M., Mattoccia, S.: Learning monocular depth estimation infusing traditional stereo knowledge. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 9799–9809. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.01003, http://openaccess.thecvf.com/content_CVPR_2019/html/Tosi_Learning_Monocular_Depth_Estimation_Infusing_Traditional_Stereo_Knowledge_CVPR_2019_paper.html

34. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: International Conference on 3D Vision (3DV) (2017)
35. Wang, C., Buenaposada, J.M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: CVPR (2018)
36. Wang, C., Buenaposada, J.M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2022–2030 (2018)
37. Wang, J., Zhang, G., Wu, Z., Li, X., Liu, L.: Self-supervised joint learning framework of depth estimation via implicit cues. arXiv:2006.09876 (2020)
38. Watson, J., Firman, M., Brostow, G.J., Turmukhambetov, D.: Self-supervised monocular depth hints. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 2162–2171. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00225, `https://doi.org/10.1109/ICCV.2019.00225`
39. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: Self-supervised multi-frame monocular depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1164–1174 (2021)
40. Wong, A., Soatto, S.: Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 5644–5653. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00579, `http://openaccess.thecvf.com/content_CVPR_2019/html/Wong_Bilateral_Cyclic_Constraint_and_Adaptive_Regularization_for_Unsupervised_Monocular_Depth_CVPR_2019_paper.html`
41. Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R.: LEGO: learning edge with geometry all at once by watching videos. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 225–234. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00031, `http://openaccess.thecvf.com/content_cvpr_2018/html/Yang_LEGO_Learning_Edge_CVPR_2018_paper.html`
42. Yin, Z., Shi, J.: GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In: CVPR (2018)
43. Zhan, H., Garg, R., Weerasekera, C.S., Li, K., Agarwal, H., Reid, I.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: CVPR (2018)
44. Zhan, H., Garg, R., Weerasekera, C.S., Li, K., Agarwal, H., Reid, I.D.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 340–349. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00043, `http://openaccess.thecvf.com/content_cvpr_2018/html/Zhan_Unsupervised_Learning_of_CVPR_2018_paper.html`
45. Zhou, J., Wang, Y., Qin, K., Zeng, W.: Unsupervised high-resolution depth learning from videos with dual networks. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 6871–6880. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00697, `https://doi.org/10.1109/ICCV.2019.00697`
46. Zhou, T., Brown, M., Snavely, N., Lowe, D.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)

47. Zou, Y., Luo, Z., Huang, J.B.: Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 36–53 (2018)