Noname manuscript No.
(will be inserted by the editor)

Self-supervised monocular depth estimation with occlusion mask and edge awareness

Shi Zhou · Miaomiao Zhu · Zhen Li · He Li · Mitsunori Mizumachi · Lifeng Zhang

Received: date / Accepted: date

Abstract Depth estimation is one of the basic and important tasks in 3D vision. Recently, many works have been done in self-supervised depth estimation based on geometric consistency between frames. However, these research works still have difficulties in ill-posed areas, such as occlusion areas and texture-less areas. This work proposed a novel self-supervised monocular depth estimation method based on occlusion mask and edge awareness to overcome these difficulties. The occlusion mask divides the image into two classes, making the training of the network more reasonable. The edge awareness loss function is designed based on the edge obtained by the traditional method, so that the method has strong robustness to various lighting conditions. Furthermore, we evaluated the proposed method on the KITTI datasets. The occlusion mask and edge awareness are both beneficial to find corresponding points in ill-posed areas.

Keywords Monocular depth estimation · Self-supervised learning · Computer vision · Edge awareness · Convolutional neural network

1 Introduction

Depth estimation is a fundamental problem in 3D computer vision. It can help perceive environments and estimate state, and has a wide range of applications in simultaneous localization and mapping (SLAM) [1], robot navigation [2], and object detection [3]. Structure from motion (SFM) [4] is a representative method, which has been successfully applied in 3D reconstruction and SLAM.

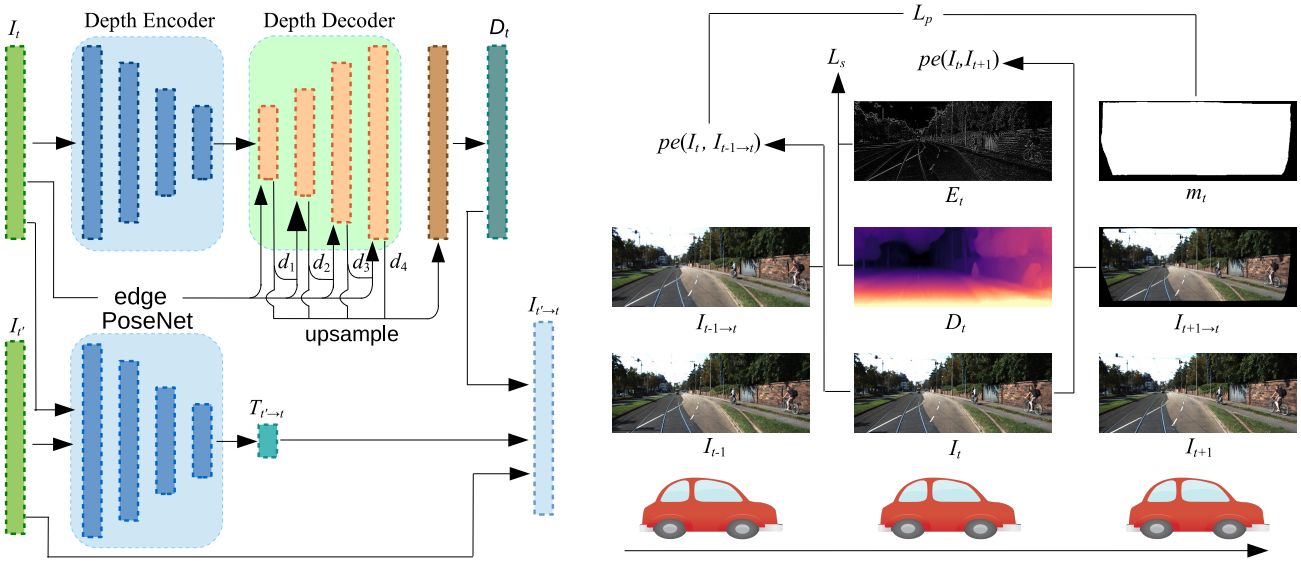
As a classical research topic for decades, many research works have been done, include the design of radar sensor and depth camera, and depth estimation method based on image information from RGB camera. However, the price of radar sensor is high, so that it is only used for some special application and cannot be widely used. Depth camera, such as Kinect camera, provide depth based on infrared. The accuracy of depth map is severely affected by the illumination. Therefore, depth camera is usually used in indoor or dark scenes, and not suitable for outdoor. According to the geometric constraints, depth can be estimated based on image information with low cost. And it can be used in both indoor and outdoor environment [5].

Depth estimation from image information always be divided into two classes: stereo matching and monocular depth estimation [6]. Stereo matching simulates the way of people understand the world with both eyes, and calculate the similarity between pixels by designing an energy function. The accuracy of depth estimation relies heavily on the energy function and the calibration accuracy of two cameras. In addition, stereo matching relies on the information of two images, which makes the hardware cost and computational cost high. However, humans can also estimate distance through one eye based on the prior understand of the world. Inspired by this, many monocular depth estimation methods are proposed with low computational cost.

With the rapid development of convolutional neural networks in image processing, many research works have been done in self-supervised monocular depth estimation based on geometric consistency between frames

S. Zhou · M. Zhu · Z. Li · M. Mizumachi · L. Zhang
Kyushu Institute of Technology, Kitakyushu, Fukuoka, Japan, 804-8550
E-mail: zhou.shi403@mail.kyutech.jp

H. Li
Northeastern University, Shenyang, Liaoning, China, 110819



(a) The architecture of DepthNet and PoseNet (b) Photometric reprojection loss and edge awareness loss

Fig. 1 Monocular depth estimation pipeline. Current frame can be re-projected based on the Depth map D_t and pose transmission $T_{t'\to t}$. Both networks are trained based on the difference between current frame and its re-projected frame and edge information of current frame.

in the past few years, instead of using hand-to-obtained ground truth. However, those methods still have difficulties in ill-posed areas, such as occlusion areas, repeat texture areas and texture-less areas. In addition, most existing methods assume that the geometric relationship between frames is same as that between stereo images, which make the learning problem harder, result in low accuracy in surrounding area. Therefore, to solve the problems described above, more research works should be devoted to self-supervised monocular depth estimation.

In this work, we proposed a novel self-supervised monocular depth estimation method for SLAM system, which can be used in outdoor environment. According to the depth map, the pose of the self-driving car and the distance to surrounding objects can be calculated, so that the car can know the location and mapping. To overcome the difficult described above, occlusion mask and edge awareness loss is proposed to better estimate depth. When the self-driving car moves forward, the surrounding information will be wiped out in the next frame. For current frame, its two adjacent frames are different. Therefore, the depth of the boundary area should only consider the information of both the current frame and the previous frame, instead of all three frames. The occlusion mask is designed to distinguish the differences and divide current frame into two classes. Edge awareness loss, which based on edge obtained by traditional Laplacian method instead of image, is proposed to strengthen the robustness of different lighting conditions. In addition, the computational cost of Laplacian edge is lower than that obtained by neural network.

2 Proposed method

In this section, we detail our self-supervised monocular depth estimation method. Firstly, making a introduction about the self-supervised training scheme which simultaneously learns depth with DepthNet and pose transmission with PoseNet. Then, according to the difference between two adjacent frames, occlusion mask is proposed to divide all pixels into two classes and better instruct the training of the network. And edge obtained by traditional Laplacian method is introduced to strengthen the robustness to different lighting conditions. At the end, an explanation is given to the final training loss.

2.1 Self-supervised depth estimation

The core concept of self-supervised monocular depth estimation is to simulate stereo vision. According to one of the stereo image, the other one can be re-projected based on the depth and transmission. In line with monocular depth estimation literature, DepthNet and PoseNet are designed to estimate depth map and pose transmission [7]. The architecture of both DepthNet and PoseNet is shown in Fig.1(a). Features of current frame

is extracted through the encoder part. Then 4 different scale depth map will be upsample to the same scale, and the final depth map will be output after the last convolutional layer.

When training the network, both current frame and its neighbor frames are inputted to the network, then the depth map of current frame and the pose transmission between adjacent frames can be estimated from the networks. According to the pose transmission and depth map, the neighbor frame can be warped into current frame. However, when testing, only DepthNet is used for depth estimation.

According to the appearance difference between current frame and its re-project frame, the parameters of DepthNet and PoseNet can be optimized by self-supervised, as shown in Fig.1(b). For current frame I_t , its depth map D can be obtained from DepthNet. The pose transmission $T_{t\to t'}$ between I_t and its neighbor frame $I'_t \in \{I_{t-1}, I_{t+1}\}$ can be obtained from PoseNet. Similarity to stereo matching, I_t can be re-projected from $I_{t'}$

$$I_{t'\to t} = proj(I_{t'}, D, T_{t\to t'}, K) \tag{1}$$

Where $proj()$ denote projector operation and K is the intrinsic metric of camera. The parameters of the network are optimized by minimizing the photometric re-projection loss between current frame I_t and its re-project frame $I_{t'\to t}$. The photometric re-projection loss L_p is a combination of L1 distance (Taxicab geometry) and structural similarity (SSIM) in pixel space

$$L_p = \frac{1}{n}\sum pe(p,q) \tag{2}$$

$$pe(p,q) = \alpha\frac{(1-\mathrm{SSIM}(p,q))}{2} + (1-\alpha)|p-q| \tag{3}$$

If the photometric re-projection loss between current frame I_t and its re-project frame $I_{t'\to t}$ is small, it means that the network can perform depth estimation well.

2.2 Occlusion mask divide all pixels into two classes

When training the network, the current frame's depth estimation usually refers to the information from the previous frame and the next frame. But when the camera moves forward, the field of view is changing each frame, all the information in the current frame can be found in the previous one, but the surrounding information will be wiped out in the next frame. Therefore, for current frame, there is a difference between the previous frame and the next frame, which is different from stereo matching.

However, existing self-supervised monocular depth estimation methods used the same way to calculate the loss for both the previous frame and the next frame. This can cause issues to pixels which located at surrounding areas and disturb the training of the two network. If different methods are used to calculate loss for the previous frame and the next frame, a better depth map will be estimated by DepthNet.

In order to remove the influence of the next frame in boundary area, the occlusion mask divides the image into two classes: occlusion pixels and non-occlusion pixels. For pixel p located in previous frame I_{t-1}, it can be re-projected into current frame

$$q = proj(p, D_t, T_{t\to t'}, K) \tag{4}$$

$$m = \begin{cases} 0 & q \in I_t \\ 1 & q \notin I_t \end{cases} \tag{5}$$

If the corresponding pixel q located in current frame, it will be classified as non-occlusion pixel, otherwise it is occlusion pixel. Because all objects in the next frame can be found in the current frame, this operation is only performed on the previous frame. For different pixels, different photometric re-projection loss is calculated to get a better depth map

$$L_p = \begin{cases} \min\limits_{t'\in\{t-1,t+1\}} pe(I_t, I_{t'}) & m = 0 \\ pe(I_t, I_{t-1}) & m = 1 \end{cases} \tag{6}$$

For any non-occlusion pixel in current frame, it can be viewed in at least one of the two adjacent frames. We take the minimum photometric re-projection loss between the two adjacent frames, so that the accuracy of occlusion surrounding areas can be improved. For any occlusion pixel in current frame, the proposed method takes the photometric re-projection loss which only considers the information of current frame and the previous frame. Without the disturb of occlusion pixels of next frame, the accuracy of both occlusion area and non-occlusion area will be improved.

2.3 Trained with edge awareness loss

For pixels in texture-less areas, their intensities are very close, even there is no difference, so the depth of these pixels will be estimated incorrectly. However, although some traditional smoothing methods can modify the wrong depth easily, they will also smooth the edge of object. Therefore, this is a difficult problem in depth estimation. Neither traditional methods nor deep learning method can solve this problem perfectly.

To penalize sudden depth changes in flat areas and encourage depth discontinuities at the edge areas of objects, previous works penalize such sudden changes in

flat areas based on image gradients, which are greatly affected by different lighting conditions. Differently, we use edge awareness based on edge, which obtained by the traditional method

$$E_t = \text{Laplacian}(I_t) \tag{7}$$

$$L_s = |\partial x D_t^*| e^{-|\partial x E_t|} + |\partial y D_t^*| e^{-|\partial y E_t|} \tag{8}$$

The edge E_t of current frame I_t is calculated by traditional Laplacian edge detection method. It is very easy to implement, and the computational cost is smaller than other methods based on convolutional neural networks. To discourage shrinking of estimated depth, mean-normalized inverse depth of I_t is considered

$$D_t^* = \frac{D_t}{\overline{D_t}} \tag{9}$$

2.4 Final training loss

For self-supervised depth estimation, we usually assume that the camera is moving forward in a static scene. However, when the camera is stop or there is an object moving in the scene, these assumptions will be broken down. Therefore, these object in the scene, or even the whole frame should be ignored.

Like some other works [8,9], a binary mask is used to divide all the pixels into two classes. If the corresponding pixels is consistent between adjacent frames, this will indicate that the camera is static, or the object is moving with same speed as camera, a stationary mask is designed to ignore these pixels

$$\mu = \begin{cases} 1 & \min\limits_{t'} pe(I_t, I_{t'}) < pe(I_t, I_{t'}) \\ 0 & \min\limits_{t'} pe(I_t, I_{t'\to t}) \geq pe(I_t, I_{t'}) \end{cases} \tag{10}$$

In order to improve the accuracy of depth map, final training loss is a combination of edge awareness loss and photometric re-projection loss

$$L = \mu L_p + \lambda L_s \tag{11}$$

In addition, since it is difficult for the studied network to finding correct correspondences at depth discontinuities areas, we increase the weight of smoothing loss in the last few epoch of training to strengthen the training of edge areas.

3 Experiments

3.1 Datasets

We use the popular KITTI datasets, which is split by Eigen *et al.* [10], for all the experiments. Before training, the same pre-processing as Zhou *et al.* [11] is performed to remove some static frames. Finally, 39810 frames are used for training, 4424 frames are used for validation, and 697 frames are used for testing. When training these networks, the same intrinsic and average focal length of cameras are used for all frames.

Table 1 Monocular depth estimation accuracy on the KITTI datasets using the Eigen split

Method	Abs Rel	Sq Rel	RMSE	RMSE log
Zhou [11]	0.183	1.595	6.709	0.270
Yang [12]	0.182	1.481	6.501	0.267
Mahjourian [13]	0.163	1.240	6.220	0.250
DDVO [14]	0.151	1.257	5.583	0.228
Ranjan [15]	0.148	1.149	5.464	0.226
EPC++ [16]	0.141	1.029	5.350	0.216
Struct2depth [17]	0141	1.026	5.291	0.215
Monodepth1 [18]	0.124	1.388	6.125	0.217
Monodepth2 [9]	0.115	0.903	4.863	0.193
Proposed	**0.112**	**0.835**	**4.812**	**0.191**

3.2 Implementation details

Both DepthNet and PoseNet are built with PyTorch. The encoder part of both networks is based on ResNet-18, so we initialize both networks with the weights pre-trained on ImageNet. The decoder part of DepthNet outputs 4 kinds of scale's depth map and we use bilinear method to upsample them to the same scale. Finally, these 4 depth maps are input to the last convolutional layer, and the required depth map is output. The output of PoseNet is a 6-dimensional vector, which is converted into a rotation matrix and a translation matrix to re-project current frame.

Adam optimizer is used to train both DepthNet and PoseNet for 20 epochs on a GTX1080 GPU with 11G of memory. The batch size is 8 and the input/output resolution is 640*192. The initial learning rate is 10^{-4}, and drop to 10^{-5} for the last 5 epochs. The initial weight λ of edge awareness loss is 10^{-3}.

3.3 Depth estimation

Here, we present and discuss the result of proposed monocular depth estimation method, and how they compare with some other similar research work.

Quantitative comparison According to the standard evaluation metrics proposed by Zhou *et al.*, we evaluate the proposed method and compared with some other self-supervised method, as show in Table.1. The depth map result is show in Fig.2. The proposed method can obtain comparable result with monodepth2, and outperform all other method. To better present the advantage of occlusion mask and edge awareness, we make an analysis below.

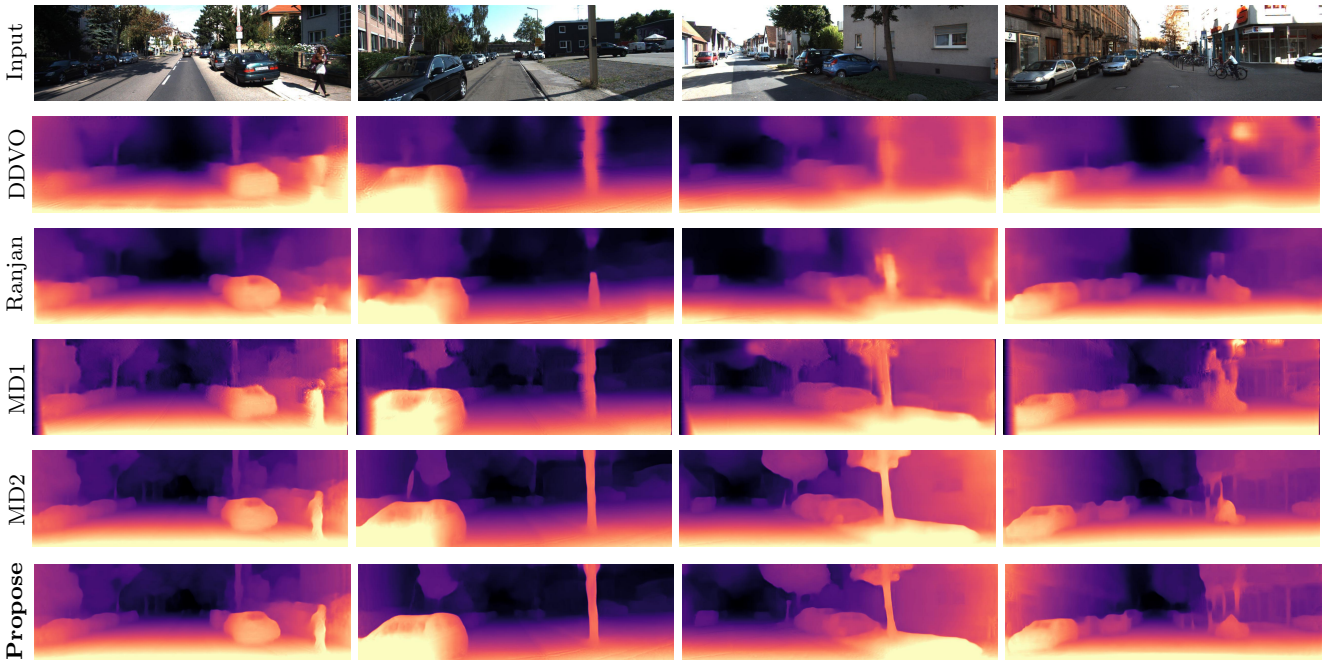


Fig. 2 Comparison of depth map on the KITTI Eigen split. The depth map of the proposed method is comparable with MD2 and better than all other method.

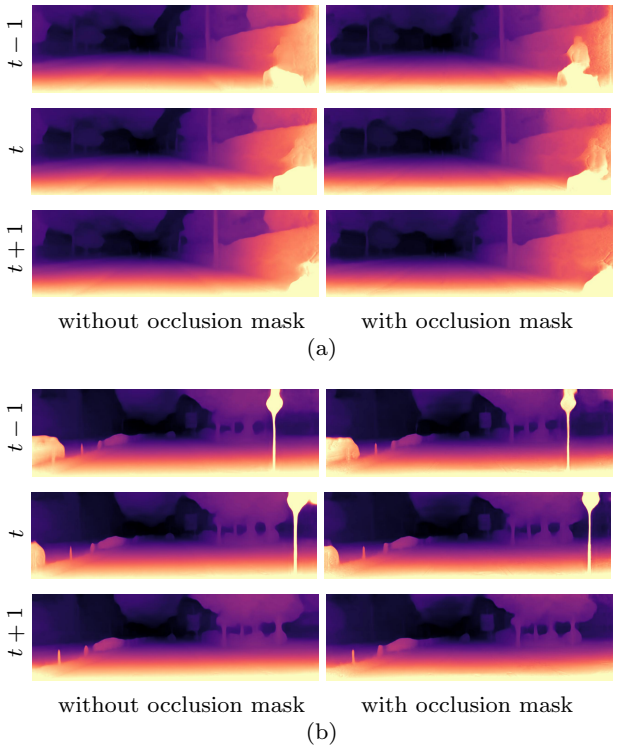


Fig. 3 Analyze the contribution of occlusion masks. With occlusion mask, better deth map can be predicted in surrounding area.

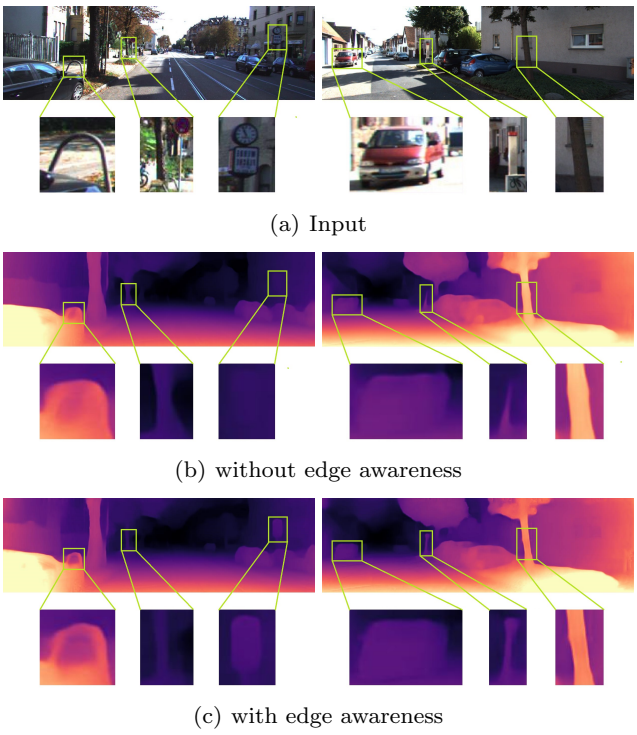


Fig. 4 Analyze the contribution of edge awareness. More accurate depth map can be predicted in the area near the edge of the object

Benefits of occlusion mask Label the area that will be warped out in the next frame with occlusion mask, so that the influence of surrounding occlusion area is removed. We analyzed the results of occlusion mask, as show in Fig.3. For the right part of depth map in frame t, which cannot be viewed in frame $t+1$ and can be

viewed in $t-1$, method without occlusion mask consider all three frame's information cannot predict a well depth map in the surrounding area. However, the proposed method with occlusion mask, which only consider the information of frame t and $t-1$ in surrounding area, and consider all three frame's information in other area, can predict a better depth map in surrounding area.

Benefits of edge awareness Encourage depth smooth in flat area and allow depth discontinuities at object boundaries, so that more semantic-meaningful edge information can be learned by the network. We analyzed the results of edge awareness loss, as show in Fig.4. Without edge awareness, the shape of object in depth map is quite different from its actual shape. For example, tree trunks become thicker or thinner, holes and small object is ignored. With edge awareness, clearer edge of objects can be estimate in the depth map.

4 Conclusion

This paper proposed a novel self-supervised method for monocular depth estimation. The result on the KITTI datasets suggests that occlusion mask and edge awareness are both beneficial to monocular depth estimation, especially the edge of object and the surrounding area of the image. It is necessary for monocular depth estimation to distinguish the relationship between the current frame and its two adjacent frames. Designing edge awareness loss with edge information obtained by traditional methods is helpful for depth estimation and has strong robustness to different lighting conditions.

References

1. Tripathi N., Sistu G., Yogamani S.: Trained trajectory based automated parking system using visual slam. arXiv preprint arXiv, pp. 1-6 (2020)
2. Xu D., Chen Y., Lin C., *et al.*: Real-time Dynamic Gesture Recognition System based on Depth Perception for Robot Navigation. In Proceedings of the IEEE Conference on Robotics and Biomimetics, pp. 689-694 (2012)
3. Song S., Xiao J.: Sliding Shapes for 3D Object Detection in Depth Images. In European conference on computer vision, pp. 634-651 (2014)
4. Garg R., Bg V. K., Carneiro G., *et al.*: Unsupervised cnn for single view depth estimation: Geometry to the rescue. European conference on computer vision (ECCV), pp. 740-756 (2016)
5. Li L., Zhang S., Yu X., *et al.*: PMSC: PatchMatch-Based Superpixel Cut for Accurate Stereo Matching. IEEE Transactions on Circuits & Systems for Video Technology, pp. 679-692 (2018)
6. Clevert D. A., Unterthiner T., Hochreiter S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv, pp. 1-14 (2015)
7. Guizilini V., Li J., Ambrus R., *et al.*: Robust Semi-Supervised Monocular Depth Estimation with Reprojected Distances. Conference on Robot Learning, pp. 503-512 (2020)
8. Guizilini V., Li J., Ambrus R., *et al.*: Robust Semi-Supervised Monocular Depth Estimation with Reprojected Distances. Conference on Robot Learning, pp. 503-512 (2020)
9. Godard C., Mac Aodha O., Firman M., *et al.*: Digging into Self-Supervised Monocular Depth Estimation. Proceedings of the IEEE international conference on computer vision, pp. 3828-3838 (2019)
10. Eigen D., Fergus R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. Proceedings of the IEEE international conference on computer vision, pp. 2650-2658 (2015)
11. Zhou T., Brown M., Snavely N., *et al.*: Unsupervised learning of depth and ego-motion from video. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1851-1858 (2017)
12. Yang Z., Wang P., Xu W., *et al.*: Unsupervised learning of geometry with edge-aware depth-normal consistency. arXiv preprint arXiv, pp. 1-8(2017)
13. Mahjourian R., Wicke M., Angelova A.: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5667-5675 (2018)
14. Wang C., Miguel Buenaposada J., Zhu R.: Learning depth from monocular videos using direct methods. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2022-2030 (2018)
15. Ranjan A., Jampani V., Balles L., Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 12240-12249 (2019)
16. Luo C., Yang Z., Wang P.: Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. IEEE transactions on pattern analysis and machine intelligence, pp. 2624-2641 (2019)
17. Casser V., Pirk S., Mahjourian R., *et al.*: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8001-8008 (2019)
18. 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)