Monocular Depth Estimation Based on Deep Learning:A Survey

1st Ruan Xiaogang
Institute of artificial intelligence and robotics, BJUT
Faculty of Information Technology, BJUT
Peking, China
adrxg@bjut.edu.cn

2nd Yan Wenjing
Institute of artificial intelligence and robotics, BJUT
Faculty of Information Technology, BJUT
Peking, China
yanwj@emails.bjut.edu.cn

3rd Huang Jing
Institute of artificial intelligence and robotics, BJUT
Faculty of Information Technology, BJUT
Peking, China
huangjing@bjut.edu.cn

4th Guo Peiyuan
Institute of artificial intelligence and robotics, BJUT
Faculty of Information Technology, BJUT
Peking, China
guopeiyuan0819@163.com

5th Guo Wei
Institute of artificial intelligence and robotics, BJUT
Faculty of Information Technology, BJUT
Peking, China
18401658535@163.com

2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9327548

***Abstract*—Monocular depth estimation relied on RGB images is an important ill posed problem in the system of computer vision. Recently, people use the method of deep learning to discuss this problem. Most of the existing monocular depth estimation algorithms relied on convolution neural network. Depth estimation based on 2D images has important applications in image segmentation, 3D object detection, robot navigation, object tracking and autonomous driving. This paper gives a brief overview of this problem, reviews, evaluates and discusses the monocular depth estimation algorithms relied on deep learning, and looks forward to the direction of further research in the face of some challenges.**

***Keywords*—monocular depth estimation, convolutional neural networks，deep learning, RGB images**

I. Introduction

Monocular depth estimation is a basic problem of computer vision. In computer vision system, three-dimensional scene information provides more possibilities for image segmentation, 3D object detection, robot navigation and other computer vision applications. Depth map, as a common expression of 3D scene information, has been widely used. The depth estimation method uses the two-dimensional information captured by the vision sensor to save the three-dimensional space information. There are a variety of devices on the market that can provide in-depth map. Nevertheless, their processing capacity, computing time, scope limitations and costs make them unsuitable for embedded devices. For example, sensors such as Kinect, which are universally used in embedded devices. Such sensors are only suitable for indoor scene and get close range depth. Similarly, laser detection and ranging (LiDAR) are usually used for three-dimensional measurements in outdoor scene. The main strengths of LiDAR is that it has high resolution and high precision. However, LiDAR are costly equipment and it is greatly affected by weather and atmosphere when working, which makes them not applicable for consumer services. Most stereo or multi view methods can accurately estimate the depth. However, they are quite time-consuming and computationally demanding.

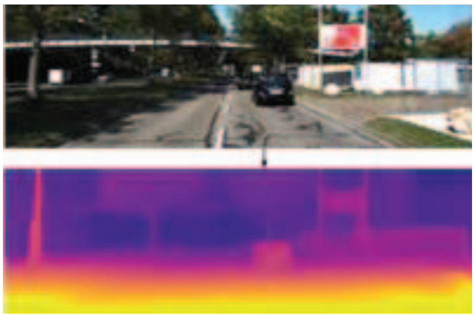

Fig. 1. Obtain scene depth prediction from RGB

The most advanced techniques have shown that monocular depth estimation algorithm may be possible solutions to these challenges. These methods have the advantages of small amount of calculation and short calculation time. They don't need to be aligned and calibrated, which is significant for multi-sensor depth estimation platforms. As a common way to express 3D scene information, depth map has been widely used. The depth map predicted from monocular RGB image is shown in Fig. 1.

This article reviews the monocular depth estimation algorithms relied on deep learning in the past five years, which are mainly subdivided into supervised depth estimation methods and self-supervised monocular depth estimation methods, and predicts the research direction of monocular depth evaluation relied on deep learning and research hotspots.

II. Survey of Monocular Depth Estimation

A. Problem Representation

Because the depth value is continuous, most of the existing methods [1] - [3] use structured regression model to model depth estimation. The parameters of these models are trained and optimized by minimizing the L2 loss function between the output and the actual depth, with the aim of outputting as close to the actual depth as possible during the assessment.

Table 1. Datasets for monocular depth estimation.

Dataset	Labeled Images	Annotation
NYU-v2[6]	1449	Depth + Segmentation
Make3D[7]	534	Depth
KITTI[8]	94k	Depth aligned with RAW

Different from the existing monocular depth estimation methods, Fu et al. [4] explored to discretize the continuous depth to multiple intervals, and transformed the depth network learning into a sequential regression problem, and proposed how to apply the sequential regression to dense prediction tasks through deep convolution neural network.

For humans, it may be difficult to measure the precise distance of an object in the surrounding environment, but we can easily estimate the range of the distance. Based on this, Cao et al. [5] divided the continuous depth data into several depth intervals, and used the classification model to model the depth estimation problem. They trained a network to output the depth range of pixels rather than the depth value.

B. *Datasets for Depth Estimation*

In order to evaluate various depth estimation algorithms, a strong benchmark is needed. In this part, we summarized three commonly used datasets for monocular depth estimation, NYU-v2, Make3D and KITTI. The common datasets are shown in Table 1.

- NYU-v2: The NYU-v2 dataset [6] samples video sequences from multiple indoor environments. Its image and depth data are from Kinect.Its characteristics are as follows: 1449 pairs of labeled stereo RGB and depth pictures we can use.
- Make3D: The dataset described in [7] includes aligned pictures and depth map. It includes indoor and outdoor scenes as well as composite objects. Its depth data comes from laser scanner.
- KITTI: The existing datasets are not large enough, and incomplete or laboratory environment, so many algorithms perform well on the datasets, but not in the real environment. The goal of KITTI is to provide a large number of real-world data sets to better measure and test the performance of the algorithm. The data acquisition platform of KITTI dataset is equipped with two gray-scale cameras, two color cameras, a velodyne 64 line 3D lidar, four optical lenses and a GPS navigation system. The whole data set is composed of 389 pairs of stereo images and optical flow maps, 39.2 km visual ranging sequence and over 200K 3D labeled object images, which are sampled and synchronized at 10Hz frequency.

C. *Evaluation Criteria*

The commonly used evaluation criteria to evaluate the good or bad of monocular depth estimation algorithms are Absolute Relative Error (AbsRel), Root Mean Square Error (RMSE), logRMSE, Square Relative Error (SqRel) and accuracy(%correct) . Generally, the smaller the error, the better, and the higher the accuracy, the better.

These index are defined as follows:

$$\mathrm{AbsRel}=\frac{1}{\mathrm{N}}\sum\frac{|\mathrm{d_i}\text{-}\mathrm{d_i^*}|}{\mathrm{d_i}} \tag{1}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{\mathrm{N}}\sum|\mathrm{d_i}\text{ -}\mathrm{d_i^*}|^2} \tag{2}$$

$$\mathrm{RMSE(log)}=\sqrt{\frac{1}{\mathrm{N}}\sum|\log\mathrm{d_i}\text{-}\log\mathrm{d_i^*}|^2} \tag{3}$$

$$\mathrm{SqRel}=\frac{1}{\mathrm{N}}\sum\frac{|\mathrm{d_i}\text{-}\mathrm{d_i^2}|^2}{\mathrm{d_i}} \tag{4}$$

$$\%\mathrm{correct:max}\left(\frac{\mathrm{d_i}}{\mathrm{d_i^*}},\frac{\mathrm{d_i^*}}{\mathrm{d_i}}\right)=\delta<\mathrm{T} \tag{5}$$

here d_i and d_i^* are the ground truth and i represents image pixel and N is the sum of all pixels. T represents the threshold, there are three values of $1.25, 1.25^2, 1.25^3$.

III. Monocular depth estimation based on deep learning

Early work commonly used hand-made feature and probability graphical model. For example, Saxena et al. [9] used a Markov random field model to estimate the absolute scale and inferred depth of different image patches. Nonparametric methods [10-13] are also used to evaluate the depth of pictures by matching the depth of the image with similar luminosity content retrieved from the database. From that time on, depth estimation began to use methods based on modern deep learning [14-16], instead of using manual feature representation [6].Table 2 summarizes some of the deep learning relied monocular depth estimation algorithms we mentioned below.

A. *Supervised Depth Estimation*

The most advanced depth estimation method based on RGB image is to train convolutional neural networks using large data sets using depth learning method [15,17,18]. Eigen et al. [14] used a two-layer convolutional neural network (CNN) to predict the whole situation and part informations respectively. Eigen and Fergus [18] combined more forecasting tasks in the same structure. Liu et al. [17] simultaneously applied deeper CNN and continuous conditional random field (CRF) to intuitively get more coherent transitions and part informations. Laina et al. [15] proposed a deep ResNet on the basis of ResNet [19], and obtained better results than [17,18]. In [20], Roy et al. explored a new depth achitecture combining random forest and convolutional neural network, and called it neural regression forest (NRF). The training results of "depth" and "parallelism" of CNN are more effective. Qi et al. [21] in order to solve the fuzzy problem in prediction, the network is trained to estimate depth and normal simultaneously. Mancini et al. [22] used a CNN combining image and optical flow image to obtain depth. The combination of early RGB images and sparse depth measurement [23, 24] also provides a new idea and achieves good results. Fu et al. [4] modeled the depth estimation network as an ordered regression problem, and used a statistical method to deal with the deep discretization problem. By using a common regression loss training network, their algorithm acquires faster synchronous convergence speed and much higher accuracy.

TABLE 2. CLASSIFICATION OF MONOCULAR DEPTH ESTIMATION METHODS BASED ON DEEP LEARNING

Method	Architecture	Category
DORN[4]	CNN	Supervised
Fastdepth[25]	Encoder-Decorder	
monoDepth[28]	CNN	Unsupervised
GeoNet[43]	CNN	
Monodepth2[38]	CNN	
Depth-VO-Feat[35]	CNN	
monoResMatch[40]	CNN	
struct2depth[46]	Encoder-Decorder	

They adopt a multi-scale network architecture to avoid excessive space sharing and simultaneously capture multiple output information. The proposed deep ordinal regression network (DORN) achieved the latest results. However, the latest monocular depth estimation method uses deep neural network with complex hierarchical structure, which is too slow to meet the requirements of robot system for rapidity. Therefore, [25] proposed an efficient coding and decoding network architecture, and further reduces the computational complexity and delay through network pruning.

B. Self-supervised Monocular Depth Estimation

In the case of no real depth information, one option is to use image reconstruction as a network constraint to train the depth estimation network. Here, the model takes a set of stereo pairs or monocular sequence images as the input of the network. The model can reduce the error of image reconstruction by training the disparity of the image and wrapping it to the nearby view.

In order to achieve the same effect as supervised depth estimation, the research of self-supervised monocular training focuses on complex depth structure, error function and imaging model.

1) Self-supervised Stereo Training: One of the input forms of self-supervised monocular depth estimation is synchronous stereo, and the output is the disparity between two images. According to the principle of binocular stereo vision, there is a linear relationship between disparity and depth. Accurate depth can be obtained by training precise disparity. In [26], a new depth discretization model for viewpoint synthesis is proposed.The continuous disparity value is obtained by training, and the method is extendedby introducing the left and right depth consistency terms[27], [28] produces better results than the supervised method. Stereo based methods have developed methods based on semi-supervised data [29,30], generative countermeasure networks(GAN) [31,32], additional consistency [33], time information [34-36] and real-time use [37]. Zhan et al. [35] proposed using binocular sequences to learn depth and visual odometer. The joint training of monocular depth and visual odometer improves the accuracy of depth prediction, because the joint training increases the constraints on the network.In terms of loss function, [28] proposed a new training loss, which strengthens the consistency between the diparity of the stereo images, and leads to higher accuracy and robustness compared with the methods in existence. Sex. Godard et al. [38] used different combinations of self-supervision for training: only monocular video (M), only stereo (S), and both (MS). A minimum reprojection loss is proposed, which can handle occlusion robustly. [20] used an adaptive geometric consistency error to improve the stability of the system to isolated points and nonLambertian regions, effectively solving the problems of occlusion and texture blur.

2) Self-supervised Monocular Training: Another input form of self-supervised depth estimation is monocular video, in which continuous time frames provide input signals. Here, two neural networks are used to evaluate the disparity and camera pose simultaneously. This method is challenging in the case of moving scene. In the earliest monocular self-supervised methods, [39] trained a depth estimation network and a pose estimation network respectively. They applied additional motion interpretation masks to solve non rigid scene motion. Nevertheless,later work of the model they provided online disabled this time, achieving excellent results. Tosi et al. [40] proposed a new depth structure monoResMatch, which aims to calculate the depth from a single input image by stereo matching the features from different angles with the input image. Their network was the first end-to-end network trained from scratch. Inspired by [41], [42] proposed a more complex motion model using multiple motion masks. However, due to insufficient experimental evaluation, we can not understand its value for the time being. [43] it also divides the motion of objects, and uses depth and optical flow to interpret motion of object. Mahjourian et al. [44] considered the 3D geometric information of the overall environment, and maintained the homogeneity of estimated 3D point cloud and object motion in continuous images. They combine a new 3D based loss with a 2D loss, based on the photometric quality of frame reconstruction, using estimated depth and adjacent object motion frames. They also include effectiveness masks to avoid punishing areas where there is no useful information. [45] proposed an unsupervised monocular learning method using structure and semantics for depth and self-motion. This method is the first to use monocular camera to learn complex dynamic scene. More effective results are obtained in challenging motion scenes.

IV. DISCUSSION

A. Comparison Analysis Based on Performance

We have summarized the result evaluation conducted in the NYU-v2 and KITTI data sets, and the specific content is shown in Table 3 and Table 4. The assessment is based on the evaluation criteria we mentioned above. Through this table, we can more clearly understand the advantages and disadvantages of the algorithm we mentioned. All figures in both tables are reported by their respective authors.

TABLE 3. EVALUATION RESULTS ON KITTI DATASET

Models	AbsRel	SqRel	RMSE	RMSE(log)	δ<1.25	δ<1.25^2	δ<1.25^3
DORN[4]	0.071	0.268	2.271	0.092	0.936	0.985	0.995
monoDepth[28]	0.114	0.898	4.935	0.206	0.861	0.949	0.976
GeoNet[43]	0.147	0.936	4.348	0.218	0.810	0.941	0.977
Monodepth2[38]	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Depth-VO-Feat[35]	0.128	0.815	4.204	0.216	0.835	0.941	0.975
monoResMatch[40]	0.096	0.673	4.351	0.184	0.890	0.961	0.981
struct2depth[46]	0.1030	0.6217	3.5546	0.1749	0.8866	0.9632	0.9846

TABLE 4. EVALUATION RESULTS ON NYU-V2 DATASET

Models	AbsRel	SqRel	RMSE	RMSE(log)	δ<1.25	δ<1.25^2	δ<1.25^3
DORN[4]	0.138	0.051	0.509	0.653	0.825	0.964	0.992
Monodepth2[38]	2.344	1.365	0.734	1.134	0.826	0.958	0.979
monoResMatch[40]	1.356	1.156	0.694	1.125	0.825	0.965	0.967

B. *Future Research Directions*

In the past few years, the performance of monocular depth estimation relied on deep learning method has improved significantly. This subject is still in the development stage and needs further research. In this part, we put forward some current research directions and further research problems in the future.

Adaptive methods can adapt to the changing environment in time or under minimum supervision, which is a promising direction in depth estimation research. At present, the most advanced depth estimation method is relied on deep learning. Although the results of these methods have been improved significantly, their computational complexity is very high. In the past, the research of efficient neural network mainly focused on image semantic segmentation and target detection. As far as we know, there are few effective designs of encoding and decoding networks for high-density image depth estimation. Therefore, a key challenge is to design an efficient codec network architecture.

Another problem to be solved is how to get more accurate depth. In general, this is affected by complex scenes, such as object occlusion, scene clutter, and object materials.

V. CONCLUSION

In this paper, the latest development of monocular depth estimation based on deep learning method is reviewed. Although these methods still have some shortcomings, they have achieved encouraging results, some of which are comparable to traditional methods in terms of accuracy. Depth estimation has developed into a new term. Deep learning and data-driven technology bring new vitality to traditional image depth estimation.

REFERENCES

[1] Li, Null Bo , et al. "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs." *Computer Vision & Pattern Recognition* IEEE Computer Society, 2015, pp. 1119–1127.

[2] Wang, Peng , et al. "Towards unified depth and semantic prediction from a single image." *Computer Vision & Pattern Recognition* IEEE, 2015, pp. 2800–2809.

[3] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

[4] Huan Fu,Mingming Gong,Chaohui Wang,Kayhan Batmanghelich and Dacheng Tao, " Deep ordinal regression network for monocular depth estimation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, p 2002-2011, December 14, 2018.

[5] Cao Yuanzhouhan, Wu Zifeng and Shen Chunhua," Estimating depth from monocular images as classification using deep fully convolutional residual networks," IEEE Transactions on Circuits and Systems for Video Technology, v 28, n 11, p 3174-3182, November 2018.

[6] Silberman, Nathan , et al. "Indoor Segmentation and Support Inference from RGBD Images." *Proceedings of the 12th European conference on Computer Vision - Volume Part V* Springer, Berlin, Heidelberg, 2012..

[7] Saxena, Ashutosh , M. Sun , and A. Y. Ng . "Make3D: Learning 3D Scene Structure from a Single Still Image." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2008).31, 824–840.

[8] Geiger A, Lenz P , Urtasun R, " Are we ready for autonomous driving? The KITTI vision benchmark suiteZ," In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.

[9] A. Saxena, S. H. Chung, and A. Y. Ng, " Learning depth from single monocular images," Advances in Neural Information Processing Systems (NIPS).2005.1161–1168.

[10] K. Karsch, C. Liu, and S. B. Kang, " Depth extraction from video using non-parametric sampling," European Conference on Computer Vision (ECCV).2012.775–788.

[11] J. Konrad, M. Wang, and P. Ishwar, " 2d-to-3d image conversion by learning depth from examples," IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, p 16-22, 2012.

[12] Karsch, Kevin , C. Liu , and S. B. Kang . "Depth Transfer: Depth Extraction from Video Using Non-Parametric Sampling." IEEE Transactions on Pattern Analysis and Machine Intelligence , 2014,no. 99, pp.1–1.

[13] Liu, Miaomiao , M. Salzmann , and X. He . "Discrete-Continuous Depth Estimation from a Single Image." *IEEE Conference on Computer Vision & Pattern Recognition* IEEE Computer Society, 2014. 716–723.

[14] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network,"Advances in Neural Information Processing Systems (NIPS). 2014. 2366–2374.

[15] Laina, Iro , et al. "Deeper Depth Prediction with Fully Convolutional Residual Networks."International Conference on 3D Vision (3DV). 2016. 239–248.

[16] Ummenhofer Benjamin, Zhou Huizhong, Uhrig Jonas, Mayer Nikolaus, Ilg Eddy, Dosovitskiy Alexey and Brox Thomas, "DeMoN:

Depth and motion network for learning monocular stereo," Conference on Computer Vision and Pattern Recognition (CVPR). 2017. 5622-5631.

[17] Liu Fayao, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image,"Conference on Computer Vision and Pattern Recognition (CVPR). 2015. 5162–5170.

[18] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," International Conference on Computer Vision .2015. 2650–2658.

[19] K. He, X. Zhang, S. Ren, and J. Sun, " Deep residual learning for image recognition," Conference on Computer Vision and Pattern Recognition . 2016.770–778.

[20] Anirban Roy and Sinisa Todorovic," Monocular depth estimation using neural regression forest," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, v 2016-December, p 5506-5514, December 9, 2016.

[21] Qi Xiaojuan, Liao Renjie, Liu Zhengzhe, Urtasun Raquel and Jia Jiaya, " GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation," Conference on Computer Vision and Pattern Recognition . 2018. 283–291.

[22] Mancini Michele, Costante Gabriele, Valigi Paolo, Ciarfuglia Thomas A, "Fast robust monocular depth estimation for obstacle detection with fully convolutional networks," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2016. 4296–4303.

[23] F. Mai and S. Karaman, "Sparse-to-dense: depth prediction from sparse depth samples and a single image," IEEE International Conference on Robotics and Automation (ICRA). 2018. 4796-4803.

[24] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera," IEEE International Conference on Robotics and Automation (ICRA). 2019. 3288-3295.

[25] Wofk Diana, Ma Fangchang, Yang Tien-Ju, Karaman Sertac and Sze Vivienne, "FastDepth: Fast Monocular Depth Estimation on Embedded Systems," International Conference on Robotics and Automation(ICRA). 2019. 6101-6108.

[26] Junyuan Xie, Ross Girshick, and Ali Farhadi, "Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks," European Conference on Computer Vision (ECCV). 2016.

[27] Ravi Garg, Vijay Kumar BG, and Ian Reid, "Unsupervised CNN for single view depth estimation: Geometry to the res-cue," European Conference on Computer Vision (ECCV). 2016.

[28] Clément Godard, Oisin Mac Aodha, and Gabriel J Bros-tow, "Unsupervised monocular depth estimation with left-right consistency," Conference on Computer Vision and Pattern Recognition (CVPR). 2017. 6602-6611.

[29] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe, "Semi-supervised deep learning for monocular depth map prediction," Conference on Computer Vision and Pattern Recognition (CVPR). 2017. 2215-2223.

[30] Luo Yue, Ren Jimmy, Lin Mude, Pang Jiahao, Sun Wenxiu, Li Hongsheng et al. "Single view stereo matching," Conference on Computer Vision and Pattern Recognition (CVPR). 2018. 155-163.

[31] Aleotti Filippo, Tosi Fabio, Poggi Matteo, Mattoccia Stefano, "Generative adversarial networks for unsupervised monocular depth prediction," European Conference on Computer Vision (ECCV). 2018. 337-354.

[32] Pilzer Andrea, Xu Dan, Puscas Mihai, Ricci Elisa, Sebe Nicu, "Unsupervised adversarial depth estimation using cycled generative networks," International Conference on 3D Vision. 2018. 587-595.

[33] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia, "Learning monocular depth estimation with unsupervised trinocular assumptions," International Conference on 3D Vision. 2018. 324-333.

[34] Li Ruihao, Wang Sen, Long Zhiqiang and Gu Dongbing, "UnDeepVO: Monocular visual odometry through unsupervised deep learning," IEEE International Conference on Robotics and Automation. 2018. 7286-7291.

[35] Zhan Huangying,Garg Ravi, Weerasekera Chamara Saroj, Li Kejie, Agarwal Harsh and Reid Ian M, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," Conference on Computer Vision and Pattern Recognition (CVPR). 2018. 340-349.

[36] Madhu, Babu V, , et al. "A Deeper Insight into the UnDEMoN: Unsupervised Deep Network for Depth and Ego-Motion Estimation." IEEE International Conference on Robotics and Automation (ICRA). 2018. 1082-1088.

[37] Poggi, Matteo , et al. "Towards Real-Time Unsupervised Monocular Depth Estimation on CPU."IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).2018. 5848-5854.

[38] Clément Godard, "Digging Into Self-Supervised Monocular Depth Estimation," Proceedings of the IEEE International Conference on Computer Vision.2019.3827-3837.

[39] Tinghui Zhou, Matthew Brown, Noah Snavely, Lowe, David G., "Unsupervised learning of depth and ego-motion from video," Conference on Computer Vision and Pattern Recognition (CVPR). 2018. 6612-6621.

[40] Tosi, Fabio , et al. "Learning monocular depth estimation infusing traditional stereo knowledge." *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE, v 2019-June, p 9791-9801, June 2019.

[41] Arunkumar Byravan and Dieter Fox, "Se3-nets: Learning rigid body motion using deep neural networks," IEEE International Conference on Robotics and Automation (ICRA). 2017. 173-180.

[42] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid,et al. "SfM-Net: Learning of structure and motion from video," arXiv,2017.

[43] Zhichao Yin and Jianping Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," Conference on Computer Vision and Pattern Recognition (CVPR). 2018. 1983-1992.

[44] Mahjourian Reza, Wicke Martin and Angelova Anelia, "Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, p 5667-5675, December 14, 2018.

[45] Casser Vincent, Pirk Soeren, Mahjourian Reza, Angelova Anelia, "Unsupervised monocular depth and ego-motion learning with structure and semantics," IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.2019.381-388.