Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 6, No. 3, 2023

A Survey of Monocular Depth Estimation based on Deep Learning

Zhen Song, Jianxing Wang

Southwest Petroleum University, Chengdu Sichuan 610500, China

Abstract: Depth information is very important for machines to perceive the environment and estimate their own state. Significant advances in robotics engineering and self-driving cars in recent decades have increased the demand for accurate depth measurements. Traditional depth estimation methods include motion structure and stereo vision matching, but these are based on the feature correspondence of multiple viewpoints, and at the same time, the predicted depth map is sparse. Depth estimation is a traditional task in computer vision that can be properly predicted by applying a variety of procedures, whereas inferring depth information from a single image is an ill-posed problem. The main objective of this paper is to provide a brief overview of the development of monocular depth estimation techniques based on deep learning. This article attempts to give an overview of supervised, unsupervised, and datasets and evaluation metrics. We conclude with a brief analysis of future developments.

Keywords: Depth Estimation; Monocular Depth Estimation; Supervised; Unsupervised.

1. Introduction

Depth estimation is crucial in computer vision, especially for understanding geometric relationships in scenes. The task consists of predicting the distance between the projection center and the 3D point corresponding to each pixel. Depth estimation has direct relevance in downstream applications such as 3D modeling, robotics, and self-driving cars. Several studies [1] have shown that depth estimation is crucial for action reasoning and execution. Currently, LiDAR, structured light depth sensors and time-of-flight sensors are used to obtain depth information [2][3]. These active depth sensors tend to be heavy, expensive, and power-hungry. At the same time, they suffer from noise and artifacts, especially from reflective or transparent surfaces. In addition, depth information can also be obtained from depth defocus [4][5], multi-view Stereo (MVS) [6][7], and obtained structure from motion (SFM) [8]. However, these methods are either time-consuming or have low depth accuracy. Therefore, depth estimation using a single image from an RGB camera is an attractive alternative to the depth estimation methods described above because of its compactness, cheapness, and low power consumption.

In the past decade, inspired by the success of deep learning on high-level vision tasks [9][10], monocular depth estimation based on supervised learning has been extensively studied. It transforms the monocular depth estimation problem into a pixel-wise regression problem and achieves impressive performance [11][12]. However, supervised learning methods rely on large labeled RGB-D datasets, which are expensive and overburdened. To avoid the need for large-scale labeled datasets, unsupervised monocular depth estimation methods have recently emerged in the literature. These methods mimic human binocular or monocular vision capabilities. Among them, the ground-truth-based loss is replaced by the image reconstruction loss [13][14].

With the rapid development of deep neural networks, deep learning-based monocular depth estimation has been extensively studied and achieved good results in accuracy. Meanwhile, a dense depth map is estimated from a single image by a deep neural network in an end-to-end manner. In order to improve the accuracy of depth estimation, different types of network frameworks, loss functions and training strategies have been proposed successively. Therefore, this paper provides an overview of current deep learning-based methods for monocular depth estimation. First, we summarize several widely used datasets and evaluation metrics in deep learning-based depth estimation. Furthermore, we review some representative methods according to different training methods: supervised and unsupervised. Finally, we discuss the challenges in monocular depth estimation and provide some ideas for future research.

2. Traditional Depth Estimation Methods

Traditional depth estimation methods mostly rely on the assumption of observations of the scene, exploiting visual cues to estimate the depth of a given scene. Traditional methods can be divided into two categories: geometric calculation methods and sensor detection methods.

Geometric algorithms recover 3D structures from two images based on geometric constraints. For example, Structure from motion (SFM) [15] is a representative method for estimating 3D structures from a series of 2D image sequences, and has been successfully applied in 3D reconstruction [16] and SLAM [17] . SFM can handle the depth of sparse features through feature correspondence and geometric constraints between image sequences, i.e., the accuracy of depth estimation largely depends on precise feature matching and high-quality image sequences. In addition, SFM suffers from the problem of monocular scale ambiguity [18]. Likewise, stereo vision matching also has the ability to recover the 3D structure of the scene from two viewpoints [19][20] viewing the scene. Stereo vision matching uses two cameras to simulate the human eye, and calculates the disparity map of the image through a cost function. Unlike the SFM process based on monocular sequences [21][22], scale information is included in the depth estimation during stereo vision matching, since the

transformation between the two cameras is calibrated in advance. Although the above geometry-based methods can efficiently compute the depth values of sparse points, these methods usually rely on image pairs or image sequences [16][20]. How to obtain dense depth maps from a single image is still a major challenge due to the lack of effective geometric solutions.

Unlike the geometric estimation method, the depth sensor detection method can directly acquire the scene without complex calculations, but it is expensive and easily affected by environmental factors. Like ultrasound and TOF for measuring depth, the known velocity of the wave is used to measure the time it takes for the transmitted pulse to reach the image sensor. RGB-D cameras and LIDAR can directly obtain the depth information of the corresponding image. RGB-D cameras have the ability to directly acquire pixel-level dense depth maps of RGB images, but their measurement range is limited and they are sensitive to outdoor sunlight [23]. Although LIDAR is widely used in the autonomous driving industry for depth measurement [24], it can only generate sparse 3D maps. Furthermore, the large size and power consumption of these depth sensors (RGB-D cameras and LIDAR) hinder their use in small robots such as drones. Due to the low cost, small size and wide application of monocular cameras, estimating dense depth maps from a single image has received more attention and has recently been well studied in an end-to-end manner based on deep learning.

Traditional depth estimation methods suffer from various limitations, including computational complexity and associated high energy consumption requirements. The current research work utilizes deep learning methods to achieve more accurate results with lower computational and energy requirements. The availability of deep learning-based methods and large-scale datasets has greatly changed monocular depth estimation methods.

3. Datasets and Evaluation Metrics in Depth Estimation

3.1. Datasets

Deep learning is a machine learning method that relies on large-scale data sets for training. A dataset of sufficient size and quality is critical to the performance of deep learning models. An ideal training set should contain diverse and representative samples to capture patterns in data and generalize them to new data. Unlike traditional machine learning methods, deep learning can automatically learn feature representations directly from raw data without manual feature engineering. Large data sets enable deep learning models to learn hierarchical feature representations, from low-level edge and texture features to high-level semantic concepts. High-quality datasets are the basis for training deep learning models for depth estimation. Containing a large amount of accurate depth annotation data can capture the spatial information of the scene and provide rich supervision signals for the network. The selection of data sets will also affect the scope of application of the method, such as indoor scenes, unmanned driving, augmented reality, etc. Therefore, in the depth estimation model, different data sets are selected or constructed according to different scenarios for model training. In this paper, several commonly used models for depth estimation are introduced.

NYU-Depth V2. The NYU Depth V2 data set [25] consists of 120K pairs of RGB and depth images. The data set uses the Microsoft Kinect sensor to collect 464 indoor scenes, and splits the indoor scenes into 249 for training. 215 for testing. The RGB image resolution in the sequence is 640 × 480 pixels. Furthermore, the diversity of lighting conditions and scene categories makes this dataset broadly representative for real-world applications [26]. This improves the transferability of research results to real environments. It is worth mentioning that the NYU-Depth V2 data set is publicly available, and researchers can compare algorithms based on a unified data set, which accelerates research progress. Overall, NYU-Depth V2 has become an important evaluation benchmark and empirical research platform in RGB-D scene analysis, and it is now a commonly used benchmark and main training dataset in supervised monocular depth estimation.

KITTI. The KITTI dataset was released by Karlsruhe Institute of Technology in 2012. [27] contains image sequences and annotation data in various driving scenarios. The resolution of the collected images is 1242 × 375 pixels. The data set has the following characteristics: First, it has a large amount of data, including as many as 150,000 high-resolution images and related 3D object detection, optical flow, parallax, semantic segmentation and other rich annotations [28]; second, the data is diverse, including Complex dynamic environments such as cities, villages, and highways; in addition, the selected driving scenarios are close to real applications and are representative [29]. In a follow-up study, Eigen et al. [10] divided the 56 scenes in the KITTI dataset into two parts, 28 for training and 28 for testing. Each scene consists of stereoscopic image pairs with a resolution of 1224 × 368.

Based on the above advantages, the KITTI dataset has been widely used in multiple important tasks such as stereo vision, object detection, motion estimation, and semantic understanding, and has promoted research progress in their respective fields [30]. A typical example is that it promotes the development of monocular depth estimation technology based on deep learning. Therefore this dataset is the most common benchmark and main training dataset in unsupervised and semi-supervised monocular depth estimation.

Make3D The Make3D dataset was originally proposed by Columbia University in 2009, which contains RGB images generated by image synthesis and corresponding depth maps [31]. The dataset has the following salient features: First, the scene is rich in themes, covering outdoor and indoor categories; second, it has accurate depth ground truth data; in addition, the image resolution is high and the depth annotation density is high. These make Make3D a high-quality dataset for early learning of 3D scene representation. But the Make3D dataset [31] only contains monocular RGB images and depth images, no stereo images. Since there are no monocular sequences or stereo image pairs in this dataset, semi-supervised and unsupervised learning methods do not use it as a training set, while supervised methods usually use it for training. Instead, it is widely used as a test set for unsupervised algorithms to evaluate the generalization ability of the network on different datasets [32].

Cityscapes The Cityscapes dataset [33] mainly focuses on the task of semantic segmentation. There are 5000 finely annotated images and 20000 coarsely annotated images in this dataset. Since this dataset does not contain true values of depth, it is only applied in the training process of several unsupervised depth estimation methods [32]. At the same

time, the dataset also includes binocular video sequences of multiple cities in different months, which can provide 22973 pairs of images for the training of the unsupervised monocular depth estimation model. The performance of the deep network is improved by pre-training the network on Cityscapes and subsequent training on other datasets.

3.2. Evaluation Metrics

In order to measure the model performance of monocular depth estimation, literature [10] provides five indicators, which are: RMSE, RMSE log, Abs Rel, Sq Rel, Accuracies, as follows:

The linear Root Mean Square Error (RMSE): defined as:

$$RMSE = \sqrt{\frac{1}{|N|}\sum_{i\in N} \| d_i - d_i^* \|^2} \quad (1)$$

The logarithm Root Mean Square Error (RMSE log): defined as:

$$RMSE\ \log = \sqrt{\frac{1}{|N|}\sum_{i\in N} \| \log(d_i) - \log(d_i^*) \|^2} \quad (2)$$

Absolute Relative Difference (Abs Rel): Defined as the average of the L1 distance between the ground truth depth and the estimated depth over all image pixels, but on the scale of the estimated depth:

$$Abs\ Rel = \frac{1}{|N|}\sum_{i\in N} \frac{|d_i - d_i^*|}{d_i^*} \quad (3)$$

Squared Relative Difference (Sq Rel): Defined as the average of the L2 distance between the ground truth depth and the estimated depth over all image pixels, but scaled by the estimated depth:

$$Sq\ Rel = \frac{1}{|N|}\sum_{i\in N} \frac{\| d_i - d_i^* \|^2}{d_i^*} \quad (4)$$

Accuracy under a threshold: is the percentage of predicted pixels whose relative error is within the threshold. The formula is expressed as:

$$Accuracies:\ \%\ of\ d_i\ s.t.\ \ \max(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}) = \delta < thr \quad (5)$$

where is the predicted depth value of pixel i, and represents the true value of depth. In addition, N represents the total number of pixels with real depth values, and thr represents the threshold value, usually 1.25, 1.25^2, 1.25^3.

4. Supervised Monocular Depth Estimation

If you follow the "checklist" your paper will conform to the requirements of the publisher and facilitate a problem-free publication process.

The supervision signal of supervised methods is based on the input ground-truth depth map, so monocular depth estimation can be regarded as a regression problem [31]. Deep neural networks aim to predict a depth map from a single image. Using the difference between the predicted depth map and the real depth map to supervise the training of the network, the L2 loss formula is as follows:

$$\mathcal{L}_2\left(d, d^*\right) = \frac{1}{N}\sum_{i}^{N} \| d - d^* \|_2^2 \quad (6)$$

Therefore, the deep network learns the depth information of the scene by approximating the ground truth. The overall flow chart is as follows:

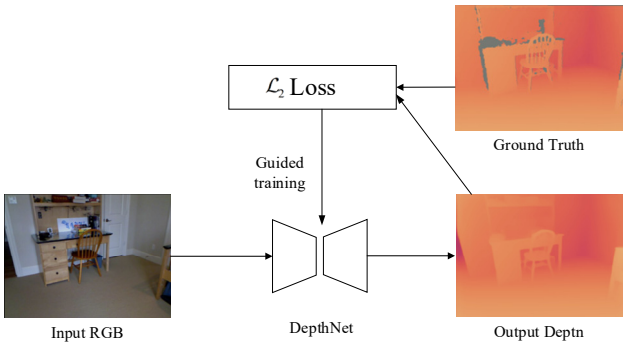

Fig 1. Supervised Basic Models

To the best of our knowledge, Eigen et al. [10] were the first to employ a CNN architecture for the task of monocular depth estimation consisting of a coarse-scale network that performs global predictions and a fine-scale network that refines predictions locally. Then, they [38] extended this method to handle three related tasks, namely depth, surface normal and semantic label prediction. Rosa et al. [34] proposed a supervised framework to estimate a continuous depth map of lidar points. The framework leverages the Hilbert map method [35] to generate a dense depth map from sparse points projected by a lidar scanner. Furthermore, the proposed framework utilizes the Fully Convolutional Residual Network (FCRN) [11] proposed by Laina et al. for depth estimation. The network is trained on dense depth images that are augmented by flipping and applying color distortions. Although the performance of this method is comparable to the state-of-the-art methods, it can only generate depth maps with a resolution of 128 × 160 pixels. More importantly, the network is biased by the output of the Hilbert map densification process, which does not represent the true depth information of the missing regions. Li et al. [36] propose a hierarchically fused dilated CNN to learn the mapping between input RGB images and corresponding depth maps. Soft weighting and inference are proposed to convert discrete depth scores to continuous depth values.

Following [11], [35], [36], Zou et al. [37] treat depth estimation as a classification problem but consider probability distributions in the training step. Their main contribution is a novel mean-variance loss, which consists of a mean loss and a variance loss. The averaging loss is used to penalize the error between the mean of the estimated depth distribution and the true value. At the same time, the variance loss is complementary to the mean loss to make the distribution sharper. The mean-variance loss is combined with a softmax loss to supervise the training of the depth estimation network.

Ladicky et al. [39] jointly predict depth maps and semantic segmentation labels through a pixel-level classifier trained by ground truth depth and semantic information, while Liu et al. [40] redefine depth prediction as a discrete Optimization problems for continuous graphical models.

From 2020 to 2021, Transformer [41] in the field of natural language processing will become popular. Researchers have introduced this attention mechanism-based network structure into the field of computer vision and achieved excellent results. In the field of depth estimation, Ranftl et al. [42] first tried to use Transformer to replace convolutional neural network as a feature extractor, and applied it to dense prediction problems. By combining global and local attention features, they achieved excellent depth estimation. result. As

the first article to introduce this structure into depth estimation, their results further triggered scholars' thinking on the Transformer structure. How to design a Transformer structure that is more suitable for depth estimation is a place that needs further exploration.

Starting from different perspectives, Bhat et al. [43] introduced Transformer as a spatial attention module into their framework. Their method further explored the potential of Transformer, a network structure, and proved that the spatial attention mechanism is beneficial to monocular Depth estimation plays an important role.

In [43], Jung et al. introduced adversarial learning to the task of monocular depth estimation. The generator consists of a global network and a refinement network, which aim to estimate global and local 3D structures from a single image. Then, a discriminator is used to distinguish the predicted and ground-truth depth maps, a form that is often used in supervised methods. The confrontation between generator and discriminator facilitates the training of min-max problem based frameworks.

5. Unsupervised Monocular Depth Estimation

During the training of unsupervised methods, instead of using expensively acquired ground truth, geometric constraints between frames are considered as supervisory signals. Another trend in monocular depth estimation is unsupervised learning, where image reconstructions of stereo image pairs or monocular video sequences are treated as supervisory signals, and depth maps are intermediate products. Xie et al. [44] used synchronized stereo images in the training phase to obtain discretized depth from soft disparity maps by minimizing the reconstruction error between the right view and the right view generated from the left view. Garg et al. [13] extended this approach to output continuous depth values, but their image formation model is not fully differentiable and thus difficult to optimize. Godard et al [32] employ bilinear samplers in a Spatial Transformation Network (STN) (Jaderberg et al [45]) for fully differentiable operations, and first introduce the left-right consistency of stereo images to train a depth estimation network. Tosi et al. [46] designed a new depth architecture for monocular depth estimation by synthesizing features from different viewpoints as input to a disparity refinement model, and proposed proxy ground-truth annotations via stereo traditional knowledge, i.e. semi- Global Matching (SGM). Wong and Soatto [47] introduced a bilateral cycle consistency constraint to enforce consistency between left and right disparities and eliminate stereo de-occlusion. Furthermore, they propose a model-driven adaptive weighting scheme to better balance data fidelity and regularization.

On the other hand, monocular video sequences are used in the training phase. For training deep prediction networks, consecutive frames in videos may have great potential as a supervised application. Camera transformation estimation (pose estimation) between consecutive frames is the main challenge of this process, which leads to additional complexity for the network. As shown in Figure 2 , Zhou et al. [14] developed a computer-based architecture to simultaneously estimate depth maps and camera poses. As input, three consecutive frames are fed to the network. Pose CNN and Depth CNN estimate relative camera pose and depth map from the first image. Bozorgtabar et al. [48] align monocular depth estimates trained on unlabeled monocular videos with the depth features of synthetic images to resolve scale ambiguities. These deep features are combined with scene depth information. Inspired by [14], Prasad and Bhowmich [50] use pole constraints to optimize joint learning of depth and ego-motion. The main idea behind the training is similar to [49]. Instead of using epipolar constraints as training labels, the authors apply it to weight pixels to guide training. Klodt and Vedaldi [51] modified [14] in the following ways. First, a structural similarity loss is introduced to enhance the brightness constancy loss. Furthermore, an explicit confidence model is incorporated into the network by predicting the likely brightness distribution of each pixel. Finally, the SFM algorithm [52] is applied to the network to provide supervisory signals for the training of the depth estimation network.

In addition to learning depth from view synthesis or minimizing photometric reconstruction errors, generative adversarial networks (GANs) also address the problem of unsupervised monocular depth estimation. GAN consists of a generator network and a discriminator network. These two networks are trained through the backpropagation algorithm, so they can work together to build an unsupervised learning model. Since there is no true depth in unsupervised learning, the discriminator distinguishes synthetic images from real images. Aleotti et al. [53] proposed the first generative adversarial network for unsupervised monocular depth estimation. A generator network is trained to infer a depth map from input images to generate warped synthetic images. Train a discriminator network to distinguish distorted images from input real images. Since the quality of the estimated depth map has an impact on the distorted synthetic image, the generator is forced to produce a more accurate depth map. Mehta et al. [54] introduced a structural adversarial training method that uses stereo view synthesis to predict dense depth maps. Given a monocular image, the generator network outputs a dense disparity map. Using the resulting disparity map, a multi-view stereo pair corresponding to the input image views is generated. A discriminator network distinguishes these reconstructed views from the real ones in the training data.

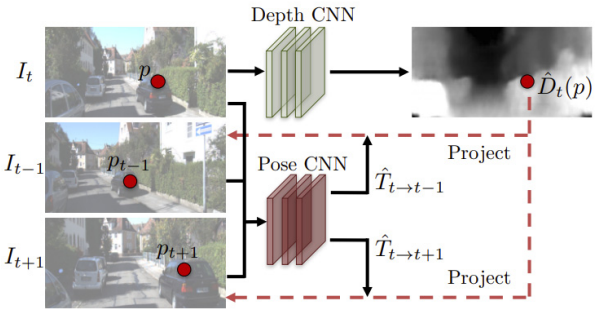

Fig 2. Developed network by Zhou et al. [14]

6. Conclusion and Future Developments

Deep learning techniques have great potential for predicting depth in monocular images. Depth prediction for monocular images can be achieved using an efficient deep learning network structure and a dataset suitable for the learning technique. This paper provides a brief overview of the contributions of this growing scientific field in deep learning-based monocular depth estimation. And the research on monocular depth estimation is reviewed from different

aspects, including training data sets, supervised and unsupervised methods, and we also introduce evaluation indicators. For the development of monocular depth estimation, from a future perspective, the architecture of deep learning models must be improved to improve the accuracy and reliability of the proposed network and reduce its inference time. At the same time, a data set composed of various scenes is also needed so that the model can learn various scene features. Therefore, how to construct a data set that satisfies deep learning has become an important research direction.

References

[1] Zhou B, Krähenbühl P, Koltun V. Does computer vision matter for action?[J]. Science Robotics, 2019, 4(30): eaaw6661.

[2] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 3354-3361.

[3] Kazmi W, Foix S, Alenyà G, et al. Indoor and outdoor depth imaging of leaves with time-of-flight and stereo vision sensors: Analysis and comparison[J]. ISPRS journal of photogrammetry and remote sensing, 2014, 88: 128-146.

[4] Wöhler C, d'Angelo P, Krüger L, et al. Monocular 3D scene reconstruction at absolute scale[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2009, 64(6): 529-540.

[5] Srinivasan P P, Garg R, Wadhwa N, et al. Aperture supervision for monocular depth estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6393-6401.

[6] Hou Y, Peng J, Hu Z, et al. Planarity constrained multi-view depth map reconstruction for urban scenes[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 139: 133-145.

[7] Mostegel C, Fraundorfer F, Bischof H. Prioritized multi-view stereo depth map generation using confidence prediction[J]. ISPRS journal of photogrammetry and remote sensing, 2018, 143: 167-180.

[8] Zeller N, Quint F, Stilla U. Depth estimation and camera calibration of a focused plenoptic camera for visual odometry[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 118: 83-100.

[9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[10] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[J]. Advances in neural information processing systems, 2014, 27.

[11] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[C]//2016 Fourth international conference on 3D vision (3DV). IEEE, 2016: 239-248.

[12] Fu H, Gong M, Wang C, et al. Deep ordinal regression network for monocular depth estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2002-2011.

[13] Garg R, Bg V K, Carneiro G, et al. Unsupervised cnn for single view depth estimation: Geometry to the rescue[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer International Publishing, 2016: 740-756.

[14] Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1851-1858.

[15] Ullman S. The interpretation of structure from motion[J]. Proceedings of the Royal Society of London. Series B. Biological Sciences, 1979, 203(1153): 405-426.

[16] Mancini F, Dubbini M, Gattelli M, et al. Using unmanned aerial vehicles (UAV) for high-resolution reconstruction of topography: The structure from motion approach on coastal environments[J]. Remote sensing, 2013, 5(12): 6880-6898.

[17] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE transactions on robotics, 2015, 31(5): 1147-1163.

[18] Szeliski R, Kang S B. Shape ambiguities in structure from motion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(5): 506-512.

[19] Zou L, Li Y. A method of stereo vision matching based on OpenCV[C]//2010 International conference on audio, language and image processing. IEEE, 2010: 185-190.

[20] Cao Z L, Yan Z H, Wang H. Summary of binocular stereo vision matching technology[J]. Journal of Chongqing University of Technology (Natural Science), 2015, 29(2): 70-75.

[21] Benosman R, Manière T, Devars J. Multidirectional stereovision sensor, calibration and scenes reconstruction [C]// Proceedings of 13th International Conference on Pattern Recognition. IEEE, 1996, 1: 161-165.

[22] Ramírez-Hernández L R, Rodríguez-Quinoñez J C, Castro-Toscano M J, et al. Improve three-dimensional point localization accuracy in stereo vision systems using a novel camera calibration method[J]. International Journal of Advanced Robotic Systems, 2020, 17(1): 1729881419896717.

[23] Tateno K, Tombari F, Laina I, et al. Cnn-slam: Real-time dense monocular slam with learned depth prediction[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 6243-6252.

[24] Yoneda K, Tehrani H, Ogawa T, et al. Lidar scan feature for localization with highly precise 3-D map[C]//2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE, 2014: 1345-1350.

[25] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgbd images[C]//Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. Springer Berlin Heidelberg, 2012: 746-760..

[26] Liu F, Shen C, Lin G, et al. Learning depth from single monocular images using deep convolutional neural fields[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(10): 2024-2039.

[27] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[28] Xie J, Kiefel M, Sun M T, et al. Semantic instance annotation of street scenes by 3d to 2d label transfer[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 3688-3697.

[29] Engel J, Schöps T, Cremers D. LSD-SLAM: Large-scale direct monocular SLAM[C]//European conference on computer vision. Cham: Springer International Publishing, 2014: 834-849.

[30] Kendall A, Cipolla R. Geometric loss functions for camera pose regression with deep learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5974-5983.

[31] Saxena A, Sun M, Ng A Y. Make3d: Learning 3d scene structure from a single still image[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 31(5): 824-840.

[32] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 270-279. 9

[33] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 3213-3223.

[34] dos Santos Rosa N, Guizilini V, Grassi V. Sparse-to-continuous: Enhancing monocular depth estimation using occupancy maps[C]//2019 19th International Conference on Advanced Robotics (ICAR). IEEE, 2019: 793-800.

[35] Ramos F, Ott L. Hilbert maps: Scalable continuous occupancy mapping with stochastic gradient descent[J]. The International Journal of Robotics Research, 2016, 35(14): 1717-1730.

[36] Li B, Dai Y, He M. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference[J]. Pattern Recognition, 2018, 83: 328-339.

[37] Zou H, Xian K, Yang J, et al. Mean-variance loss for monocular depth estimation[C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 1760-1764.

[38] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2650-2658.

[39] Ladicky L, Shi J, Pollefeys M. Pulling things out of perspective[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 89-96.

[40] Liu M, Salzmann M, He X. Discrete-continuous depth estimation from a single image[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 716-723.

[41] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[42] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction [C]// Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12179-12188.

[43] Bhat S F, Alhashim I, Wonka P. Adabins: Depth estimation using adaptive bins[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4009-4018.

[44] Xie J, Girshick R, Farhadi A. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks [C]// Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016: 842-857.

[45] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[J]. Advances in neural information processing systems, 2015, 28.

[46] Tosi F, Aleotti F, Poggi M, et al. Learning monocular depth estimation infusing traditional stereo knowledge[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9799-9809.

[47] Wong A, Soatto S. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5644-5653.

[48] Bozorgtabar B, Rad M S, Mahapatra D, et al. Syndemo: Synergistic deep feature alignment for joint learning of depth and ego-motion[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4210-4219.

[49] Mayer N, Ilg E, Hausser P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4040-4048.

[50] Prasad V, Bhowmick B. Sfmlearner++: Learning monocular depth & ego-motion using meaningful geometric constraints [C]// 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019: 2087-2096.

[51] Klodt M, Vedaldi A. Supervising the new with the old: learning sfm from sfm[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 698-713.

[52] Mur-Artal R, Tardós J D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras[J]. IEEE transactions on robotics, 2017, 33(5): 1255-1262.

[53] Aleotti F, Tosi F, Poggi M, et al. Generative adversarial networks for unsupervised monocular depth prediction[C]// Proceedings of the European conference on computer vision (ECCV) workshops. 2018: 0-0.

[54] Mehta I, Sakurikar P, Narayanan P J. Structured adversarial training for unsupervised monocular depth estimation [C]// 2018 International Conference on 3D Vision (3DV). IEEE, 2018: 314-323.

[55] Zhao C, Sun Q, Zhang C, et al. Monocular Depth Estimation Based on Deep Learning: An Overview[J]. Science China Technological Sciences, 2020, 63(9): 1612-1627.

[56] Masoumian A, Rashwan H A, Cristiano J, et al. Monocular Depth Estimation Using Deep Learning: A Review[J]. Sensors, 2022, 22(14): 5353.