SCIENCE CHINA
Technological Sciences

Special Topic: Industrial Artificial Intelligence
• Review •

September 2020 Vol. 63 No. 9: 1612–1627
https://doi.org/10.1007/s11431-020-1582-8

# Monocular depth estimation based on deep learning: An overview

ZHAO ChaoQiang, SUN QiYu, ZHANG ChongZhen, TANG Yang* & QIAN Feng

*Key Laboratory of Advanced Control and Optimization for Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China*

Received February 27, 2020; accepted March 25, 2020; published online June 10, 2020

Depth information is important for autonomous systems to perceive environments and estimate their own state. Traditional depth estimation methods, like structure from motion and stereo vision matching, are built on feature correspondences of multiple viewpoints. Meanwhile, the predicted depth maps are sparse. Inferring depth information from a single image (monocular depth estimation) is an ill-posed problem. With the rapid development of deep neural networks, monocular depth estimation based on deep learning has been widely studied recently and achieved promising performance in accuracy. Meanwhile, dense depth maps are estimated from single images by deep neural networks in an end-to-end manner. In order to improve the accuracy of depth estimation, different kinds of network frameworks, loss functions and training strategies are proposed subsequently. Therefore, we survey the current monocular depth estimation methods based on deep learning in this review. Initially, we conclude several widely used datasets and evaluation indicators in deep learning-based depth estimation. Furthermore, we review some representative existing methods according to different training manners: supervised, unsupervised and semi-supervised. Finally, we discuss the challenges and provide some ideas for future researches in monocular depth estimation.

**autonomous systems, monocular depth estimation, deep learning, unsupervised learning**

**Citation:** Zhao C Q, Sun Q Y, Zhang C Z, et al. Monocular depth estimation based on deep learning: An overview. Sci China Tech Sci, 2020, 63: 1612–1627, https://doi.org/10.1007/s11431-020-1582-8

## 1 Introduction

Estimating depth information from images is one of the basic and important tasks in computer vision, which can be widely used in simultaneous localization and mapping (SLAM) [1], navigation [2], object detection [3] and semantic segmentation [4], etc.

**Geometry-based methods** Recovering 3D structures from a couple of images based on geometric constraints is a popular way to perceive depth, and it has been widely investigated in recent forty years. Structure from motion (SfM) [5] is a representative method for estimating 3D structures from a series of 2D image sequences and is applied in 3D reconstruction [6] and SLAM [7] successfully. The depth of sparse features can be handled by SfM through feature correspondences and geometric constraints between image sequences, i.e., the accuracy of depth estimation relies heavily on the exact feature matching and high-quality image sequences. Furthermore, SfM suffers from monocular scale ambiguity [8]. Similarly, stereo vision matching also has the ability to recover 3D structures of a scene by observing the scene from two viewpoints [9, 10]. Stereo vision matching simulates the way of human eyes by two cameras, and the disparity maps of images are calculated through a cost function. Since the transformation between two cameras is calibrated in advance, the scale information is included in depth estimation during the stereo vision matching process, which is different from the SfM process based on monocular sequences [11, 12].

Although the above geometry-based methods can efficiently calculate the depth values of sparse points, these

*Corresponding author (email: yangtang@ecust.edu.cn; tangtany@gmail.com)

 

methods usually depend on image pairs or image sequences [6, 10]. How to get the dense depth map from a single image is still a significant challenge because of lack of effective geometric solutions.

**Sensor-based methods** Depth sensors, like RGB-D cameras and LIDAR, are able to get the depth information of the corresponding image directly. RGB-D cameras have the ability to get the pixel-level dense depth map of RGB image directly, but they suffer from the limited measurement range and outdoor sunlight sensitivity [13]. Although LIDAR is widely used in unmanned driving industry for depth measurement [14], it can only generate the sparse 3D map. Besides, the large size and power consumption of these depth sensors (RGB-D cameras and LIDAR) affect their applications to small robotics, like drones. Due to the low cost, small size and wide applications of monocular cameras, estimating the dense depth map from a single image has received more attention, and it has been well researched recently based on deep learning in an end-to-end manner.

**Deep learning-based methods** With the rapid development in deep learning, deep neural networks show their outstanding performance on image processing, like image classification [15], objective detection [16] and semantic segmentation [17], etc., and related well-written overviews can be found in refs. [18–21]. Besides, recent developments have shown that the pixel-level depth map can be recovered from a single image in an end-to-end manner based on deep learning [22]. A variety of neural networks have manifested their effectiveness to address the monocular depth estimation, such as convolutional neural networks (CNNs) [23], recurrent neural networks (RNNs) [24], variational auto-encoders (VAEs) [25] and generative adversarial networks (GANs) [26]. The main goal of this overview is to provide an intuitive understanding of mainstream algorithms that have made significant contributions to monocular depth estimation. We review some related works in monocular depth estimation from the aspect of learning methods, including the loss function and network framework design, which is different from our previous review [21]. Some examples of monocular depth estimation based on deep learning are shown in Figure 1.

This survey is organized in the following way. Section 2 introduces some widely used datasets and evaluation indicators in monocular depth estimation. Section 3 reviews some representative depth estimation methods based on deep learning according to different training modes. We also conclude some novel frameworks that can effectively improve network performance. Section 4 summarizes the current challenges and promising directions to research. Section 5 concludes this review.

## 2 Datasets and evaluation indicators in depth estimation

### 2.1 Datasets

**KITTI** The KITTI dataset [32] is the largest and most commonly used dataset for the sub-tasks in computer vision, like optical flow [33], visual odometry [34], depth [35], object detection [36], semantic segmentation [37] and tracking [38], etc. It is also the commonest benchmark and the primary training dataset in the unsupervised and semi-supervised monocular depth estimation. The real images from "city", "residential" and "road" categories are collected in the KITTI dataset, and the 56 scenes in the KITTI dataset

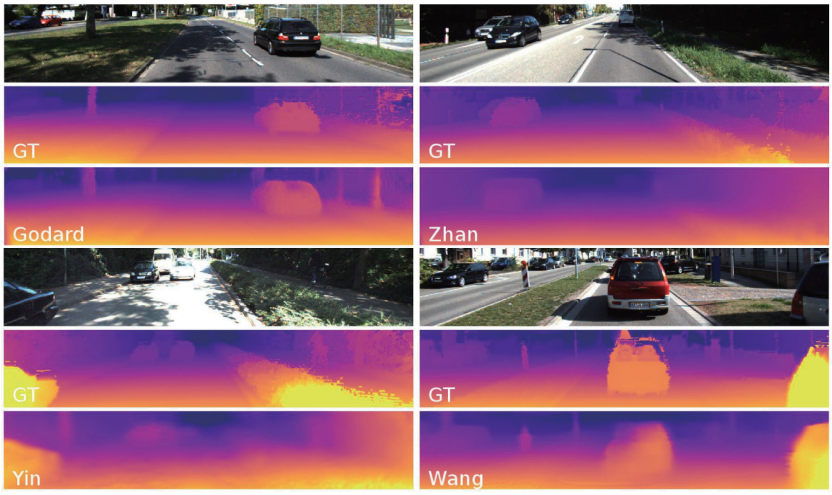


**Figure 1** (Color online) An example of monocular depth estimation. GT refers to the ground truth of depth map. The depth maps are predicted from deep neural networks proposed by Godard et al. [27], Zhan et al. [28], Yin and Shi [29], and Wang et al. [30]. The results are taken from ref. [31]. As shown in these figures, the 3D structures of objects, like trees, streets and cars, can be effectively perceived from single images by deep depth networks.

are divided into two parts, 28 ones for training and the other 28 ones for testing, by Eigen et al. [35]. Each scene consists of stereo image pairs with a resolution of 1224×368. The corresponding depth of every RGB image is sampled in a sparse way by a rotating LIDAR sensor. Since the dataset also provides the ground truth of pose for 11 odometry sequences, it is also widely used to evaluate deep learning-based visual odometry (VO) algorithms [39,40].

**NYU Depth** The NYU Depth dataset [41] focuses on indoor environments, and there are 464 indoor scenes in this dataset. Different from the KITTI dataset, which collects ground truth with LIDAR, the NYU Depth dataset takes monocular video sequences of scenes and the ground truth of depth by an RGB-D camera. It is the common benchmark and the primary training dataset in the supervised monocular depth estimation. These indoor scenes are split into 249 ones for training and 215 ones for testing. The resolution of the RGB images in sequences is 640×480, and they are also down-sampled by half during experiments. Because of the different variable frame rates between RGB camera and depth camera, it is not a one-to-one correspondence between depth maps and RGB images. In order to align the depth maps and the RGB images, each depth map is associated with the closest RGB image at first. Then, with the geometrical relationship provided by the dataset, the camera projections are used to align depth and RGB pairs. Since the projection is discrete, not all pixels have a corresponding depth value, and thus the pixels with no depth value are masked off during the experiments.

**Cityscapes** The Cityscapes dataset [42] mainly focuses on semantic segmentation tasks [37]. There are 5000 images with fine annotations and 20000 images with coarse annotations in this dataset. Meanwhile, this dataset consists of a set of stereo video sequences, which are collected from 50 cities for several months. Since this dataset does not contain the ground truth of depth, it is only applied to the training process of several unsupervised depth estimation methods [27,29]. The performance of depth networks is improved by pre-training the networks on the Cityscapes, and the experiments in refs. [27,29,43,44] have proved the effectiveness of this joint training method. The training data consists of 22973 stereo image pairs with a resolution of 1024×2048 collected from different cities.

**Make3D** The Make3D dataset [45] only consists of monocular RGB as well as depth images and does not have stereo images, which is different from the above datasets. Since there are no monocular sequences or stereo image pairs in this dataset, semi-supervised and unsupervised learning methods do not use it as the training set, while supervised methods usually adopt it for training. Instead, it is widely used as a testing set of unsupervised algorithms to evaluate the generalization ability of networks on different datasets [27].

### 2.2 Evaluation metrics

In order to evaluate and compare the performance of various depth estimation networks, a commonly accepted evaluation method is proposed in ref. [35] with five evaluation indicators: RMSE, RMSE log, Abs Rel, Sq Rel, Accuracies. These indicators are formulated as

$$\bullet\ \text{RMSE} = \sqrt{\frac{1}{|N|}\sum_{i\in N} \| d_i - d_i^* \|^2},$$

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

$$\bullet\ \text{Abs Rel} = \frac{1}{|N|}\sum_{i\in N} \frac{| d_i - d_i^* |}{d_i^*},$$

$$\bullet\ \text{Sq Rel} = \frac{1}{|N|}\sum_{i\in N} \frac{\| d_i - d_i^* \|^2}{d_i^*},$$

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

where $d_i$ is the predicted depth value of pixel $i$, and $d_i^*$ stands for the ground truth of depth. Besides, $N$ denotes the total number of pixels with real-depth values, and $thr$ denotes the threshold.

## 3 Monocular depth estimation based on deep learning

Since humans can use priori information of the world, it is capable for them to perceive the depth information from a single image. Inspired by this, previous works achieve single-image depth estimation by combining some prior information, like the relationship between some geometric structures (sky, ground, buildings) [46]. With the convincing performance in image processing, CNNs have also demonstrated a strong ability to accurately estimate dense depth maps from single images [35]. Ref. [47] investigated which kind of cues the depth networks should exploit for monocular depth estimation based on the four published methods (MonoDepth [27], SfMLearner [43], Semodepth [48] and LKVOLearner [30]).

Deep neural networks can be regarded as a black box, and the depth network will learn some structural information for depth inference with the help of supervised signals. However, one of the biggest challenges of deep learning is the lack of enough datasets with ground truth, which is expensive to acquire. Therefore, in this section, we review the monocular

depth estimation methods from the aspect of using ground truth: supervised methods [49], unsupervised methods [50] and semi-supervised methods [48]. Although the training processes of the unsupervised and semi-supervised methods rely on monocular videos or stereo image pairs, the trained depth networks predict depth maps from single images during the testing. We summarize the existing methods from the aspect of their training data, supervised signals and contributions in Table 1. We also collect the quantitative results of the unsupervised and semi-supervised algorithms evaluated on the KITTI dataset in Table 2.

### 3.1 Supervised monocular depth estimation

**A basic model for supervised methods** The supervisory signal of supervised methods is based on the ground truth of depth maps, so that monocular depth estimation can be regarded as a regressive problem [35]. The deep neural networks are designed to predict depth maps from single images. The differences between the predicted and real depth maps are utilized to supervise the training of networks, $\mathcal{L}_2$ loss:

$$\mathcal{L}_2(d, d^*) = \frac{1}{N}\sum_i^N \|d - d^*\|_2^2. \tag{1}$$

Therefore, depth networks learn the depth information of scenes by approximating the ground truth.

**Methods based on different architectures and loss functions** To the best of our knowledge, Eigen et al. [35] solved the monocular depth estimation problem by CNNs. The proposed architecture which is composed of two-component stacks (the global coarse-scale network and the local fine-scale network) is designed in ref. [35] to predict the depth map from a single image in an end-to-end way. During the training process, they use the ground truth of depth $d^*$ as the supervised signals, and the depth network predicts the log depth as $\log(d)$. The training loss function is set as

$$\mathcal{L}(d, d^*) = \frac{1}{N}\sum_i^N y_i^2 - \frac{\lambda}{N^2}\Big(\sum_i^N y_i\Big)^2, \tag{2}$$

where $y_i^2 = \log(d) - \log(d^*)$. $\lambda$ refers to the balance factor and is set to 0.5. The coarse-scale network is trained at first, and then the fine-scale network is trained to refine the results by fixing the parameters of the coarse-scale network. The experiments show that the fine-scale network is effective to refine the depth map estimated by the coarse-scale network. Eigen et al. [55] proposed a general multi-scale framework capable of dealing with the tasks such as depth map estimation, surface normal estimation, and semantic label prediction from a single image. For depth estimation, based on eq. (2), an additional loss function is proposed to promote the local structural consistency:

$$\mathcal{L}_s = \frac{1}{N}\sum_i^n \Big[(\nabla_x D_i)^2 + (\nabla_y D_i)^2\Big], \tag{3}$$

where $D_i = \log(d_i) - \log(d_i^*)$, and $\nabla$ is the vector differential operator. This function calculates the gradients of the difference between the predicted depth and the ground truth in the horizontal and vertical directions. Similarly, considering that optical flow is successfully solved by CNN through supervised learning, Mayer et al. [33] extended the optical flow networks to disparity and scene flow estimation. A fully CNN framework for monocular depth estimation is proposed in ref. [54], and then the proposed framework jointly optimizes the intrinsic factorization to recover the input image. Inspired by the outstanding performance of ResNet [92], Laina et al. [59] introduced residual learning to learn the mapping relation between depth maps and single images, therefore their network is deeper than previous works in depth estimation with higher accuracy. Besides, the fully-connected layers in ResNet are replaced by up-sampling blocks to improve the resolution of the predicted depth map. During the training process, they use the reverse Huber (Berhu) [93] as the supervised signal of depth network, which was also used in ref. [61] and achieves a better result than $\mathcal{L}_2$ loss (eq. (1)), and Berhu loss is

$$\mathcal{L}_{\text{Berhu}}(d, d^*) = \begin{cases} |d - d^*|, & \text{if } |d - d^*| \le c, \\ \dfrac{(d - d^*)^2 + c^2}{2c}, & \text{if } |d - d^*| > c, \end{cases} \tag{4}$$

where $c$ is a threshold and set to $\frac{1}{5}\max_i(|d - d^*|)$. If $|x| < c$, the Berhu loss is equal to $\mathcal{L}_1$ norm, and the Berhu loss is equal to $\mathcal{L}_2$ when $|x|$ is outside this range. Because of the deeper fully convolutional network and the improved loss function, this work achieves a better result than the previous works [33,54] with fewer parameters and training data.

Mancini et al. [94] focused on the application of depth estimation in obstacle detection. Instead of predicting the depth from a single image, the proposed fully CNN framework in ref. [94] uses both monocular image and corresponding optical flow to estimate an accurate depth map. Chen et al. [66] tackled the challenge of perceiving the single-image depth estimation in the wild by exploring a novel algorithm. Different from using the ground truth of depth as the supervised signal, their networks are trained by the relative depth annotations. The variant of Inception Module [56] is also utilized in their framework to make the network deeper. Since the monocular view contains few geometric details, Kendall et al. [49] designed a deep learning framework to learn the structure of

**Table 1** A summary of deep learning-based monocular depth estimation. "Mono." refers to "Monocular", and "multi-tasks" means that in addition to pose and depth estimation, there are other tasks that are jointly trained in the framework, such as semantic segmentation, motion segmentation, optical flow, camera intrinsic, objects motion, surface normal, etc.

| Methods | Years | Training set | Supervised (Sup) manner | | | Main contributions |
|---|---|---|---|---|---|---|
| | | | Sup | Semi-sup | Unsup | |
| Eigen et al. [35] | 2014 | RGB + Depth | √ | | | CNNs |
| Li et al. [51] | 2015 | RGB + Depth | √ | | | Hierarchical CRFs |
| Liu et al. [52] | 2015 | RGB + Depth | √ | | | Continuous CRF |
| Wang et al. [53] | 2015 | RGB + Depth | √ | | | Multi-task, hierarchical CRFs |
| Shelhamer et al. [54] | 2015 | RGB + Depth | √ | | | Fully CNNs |
| Eigen et al. [55] | 2015 | RGB + Depth | √ | | | Multi-task |
| Szegedy et al. [56] | 2015 | RGB + Depth | √ | | | Inception Module |
| Mousavian et al. [57] | 2016 | RGB + Depth | √ | | | Multi-task |
| Roy et al. [58] | 2016 | RGB + Depth | √ | | | RFs |
| Mayer et al. [33] | 2016 | RGB + Disparity | √ | | | Multi-task |
| Laina et al. [59] | 2016 | RGB + Depth | √ | | | Residual learning |
| Jung et al. [60] | 2017 | RGB + Depth | √ | | | Adversarial learning |
| Kendall et al. [49] | 2017 | Stereo images + Disparity | √ | | | Disparity loss |
| Zhang et al. [61] | 2018 | RGB + Depth | √ | | | Task-attentional, BerHu loss |
| Xu et al. [62] | 2018 | RGB + Depth | √ | | | Continuous CRF, structured attention |
| Lore et al. [63] | 2018 | RGB + Depth | √ | | | Conditional GAN |
| Fu et al. [64] | 2018 | RGB + Depth | √ | | | Ordinal regression |
| Facil et al. [22] | 2019 | RGB + Depth | √ | | | Transferability |
| Wofk et al. [65] | 2019 | RGB + Depth | √ | | | Lightweight network |
| Garg et al. [23] | 2016 | Stereo images | | √ | | Stereo framework |
| Chen et al. [66] | 2016 | RGB + Relative depth annotations | | √ | | The wild scene |
| Godard et al. [27] | 2017 | Stereo images | | √ | | Left-right consistency loss |
| Kuznietsov et al. [48] | 2017 | Stereo images + LiDAR | | √ | | Direct image alignment loss |
| Poggi et al. [67] | 2018 | Stereo images | | √ | | Trinocular assumption |
| Ramirez et al. [68] | 2018 | Stereo images + Semantic Label | | √ | | Multi-task |
| Aleotti et al. [26] | 2018 | Stereo images | | √ | | GAN |
| Pilzer et al. [69] | 2018 | Stereo images | | √ | | Cycled generative network |
| Luo et al. [70] | 2018 | Stereo images | | √ | | Stereo matching |
| He et al. [71] | 2018 | Stereo images +LIDAR | | √ | | Weak-supervised framework |
| Pilzer et al. [72] | 2019 | Stereo images | | √ | | Knowledge distillation |
| Tosi et al. [73] | 2019 | Stereo images | | √ | | Stereo matching |
| Chen et al. [74] | 2019 | Stereo images | | √ | | Multi-task |
| Fei et al. [31] | 2019 | Stereo images + IMU + Semantic Label | | √ | | Multi-task, physical information |
| Feng et al. [75] | 2019 | Stereo images | | √ | | Stacked-GAN |
| Wang et al. [30] | 2018 | Mono. sequences | | √ | | Direct VO |
| Zhan et al. [28] | 2018 | Stereo sequences | | √ | | Deep feature reconstruction |
| Li et al. [76] | 2018 | Stereo sequences | | √ | | Absolute scale recovery |
| Wang et al. [77] | 2019 | Stereo sequences | | √ | | Multi-task |
| Zhao et al. [78] | 2019 | Stereo images + Synthesized GT | | √ | | Domain adaptation, cycle GAN |
| Wu et al. [79] | 2019 | Mono. sequences + LIDAR | | √ | | Attention mechanism, GAN |
| Zhou et al. [43] | 2017 | Mono. sequences | | | √ | Monocular framework, mask network |
| Vijayanarasimhan et al. [80] | 2017 | Mono. sequences | | | √ | Multi-task |
| Yang et al. [81] | 2017 | Mono. sequences | | | √ | Multi-task |
| Mahjourian et al. [50] | 2018 | Mono. sequences | | | √ | ICP loss |
| Yin and Shi [29] | 2018 | Mono. sequences | | | √ | Multi-task |
| Zou et al. [82] | 2018 | Mono. sequences | | | √ | Multi-task |
| Kumar et al. [83] | 2018 | Mono. sequences | | | √ | GAN |
| Sun et al. [84] | 2019 | Mono. sequences | | | √ | Cycle-consistent loss |
| Wang et al. [85] | 2019 | Mono. sequences | | | √ | Geometry mask |
| Bian et al. [44] | 2019 | Mono. sequences | | | √ | Scale-consistency |
| Casser et al. [86] | 2019 | Mono. sequences | | | √ | Multi-task |
| Ranjan et al. [87] | 2019 | Mono. sequences | | | √ | Multi-task |
| Chen et al. [88] | 2019 | Mono. sequences | | | √ | Multi-task |
| Gordon et al. [89] | 2019 | Mono. sequences | | | √ | Multi-task |
| Li et al. [90] | 2019 | Mono. sequences | | | √ | GAN, LSTM, mask |
| Almalioglu et al. [91] | 2019 | Mono. sequences | | | √ | GAN, LSTM |

**Table 2** Monocular depth results of semi-supervised and unsupervised methods on the KITTI dataset [32]. "Cap" stands for the upper limit of predicted depth, and "sup" refers to "supervised". We show the best results in bold

| Method | Year | Training pattern | Cap (m) | Lower is better | | | | Accuracy: higher is better | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Abs Rel | Sq Rel | RMSE | RMSE log | $\delta < 1.25^1$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
| Garg et al. [23] L12 Aug8 × cap 50 m | 2016 | Semi-sup | 50 | 0.169 | 1.080 | 5.104 | 0.273 | 0.740 | 0.904 | 0.962 |
| Godard et al. [27] | 2017 | Semi-sup | 80 | 0.148 | 1.344 | 5.927 | 0.247 | 0.862 | 0.960 | 0.964 |
| Kuznietsov et al. [48] | 2017 | Semi-sup | 80 | 0.113 | 0.741 | 4.621 | 0.189 | 0.803 | 0.922 | 0.986 |
| Poggi et al. [67] | 2018 | Semi-sup | 80 | 0.126 | 0.961 | 5.205 | 0.220 | 0.835 | 0.941 | 0.974 |
| Ramirez et al. [68] | 2018 | Semi-sup | 80 | 0.143 | 2.161 | 6.526 | 0.222 | 0.850 | 0.939 | 0.972 |
| Aleotti et al. [26] | 2018 | Semi-sup | 80 | 0.119 | 1.239 | 5.998 | 0.212 | 0.846 | 0.940 | 0.976 |
| Pilzer et al. [69] | 2018 | Semi-sup | 80 | 0.152 | 1.388 | 6.016 | 0.247 | 0.789 | 0.918 | 0.965 |
| Luo et al. [70] | 2018 | Semi-sup | 80 | 0.094 | **0.626** | 4.252 | 0.177 | 0.891 | 0.965 | 0.984 |
| He et al. [71] | 2018 | Semi-sup | 80 | 0.110 | 1.085 | 5.628 | 0.199 | 0.855 | 0.949 | 0.981 |
| Pilzer et al. [72] | 2019 | Semi-sup | 80 | 0.098 | 0.831 | 4.656 | 0.202 | 0.882 | 0.948 | 0.973 |
| Tosi et al. [73] | 2019 | Semi-sup | 80 | 0.111 | 0.867 | 4.714 | 0.199 | 0.864 | 0.954 | 0.979 |
| Chen et al. [74] | 2019 | Semi-sup | 80 | 0.118 | 0.905 | 5.096 | 0.211 | 0.839 | 0.945 | 0.977 |
| Feng et al. [75] | 2019 | Semi-sup | 80 | **0.065** | 0.673 | **4.003** | **0.136** | **0.944** | **0.979** | **0.991** |
| Zhou et al. [43] | 2017 | Unsup | 80 | 0.208 | 1.768 | 6.865 | 0.283 | 0.678 | 0.885 | 0.957 |
| Yang et al. [81] | 2017 | Unsup | 80 | 0.182 | 1.481 | 6.501 | 0.267 | 0.725 | 0.906 | 0.963 |
| Mahjourian et al. [50] | 2018 | Unsup | 80 | 0.163 | 1.240 | 6.221 | 0.250 | 0.762 | 0.916 | 0.968 |
| Yin and Shi [29] | 2018 | Unsup | 80 | 0.155 | 1.296 | 5.857 | 0.233 | 0.793 | 0.931 | 0.973 |
| Zou et al. [82] | 2018 | Unsup | 80 | 0.150 | 1.124 | 5.507 | 0.223 | 0.806 | 0.933 | 0.973 |
| Wang et al. [85] | 2019 | Unsup | 80 | 0.158 | 1.277 | 5.858 | 0.233 | 0.785 | 0.929 | 0.973 |
| Bian et al. [44] | 2019 | Unsup | 80 | 0.137 | 1.089 | 5.439 | 0.217 | 0.830 | 0.942 | 0.975 |
| Casser et al. [86] | 2019 | Unsup | 80 | 0.109 | 0.825 | **4.750** | **0.187** | 0.874 | **0.958** | **0.983** |
| Ranjan et al. [87] | 2019 | Unsup | 80 | 0.140 | 1.070 | 5.326 | 0.217 | 0.826 | 0.941 | 0.975 |
| Chen et al. [88] | 2019 | Unsup | 80 | **0.100** | **0.811** | 4.806 | 0.189 | **0.875** | **0.958** | 0.982 |
| Li et al. [90] | 2019 | Unsup | 80 | 0.150 | 1.127 | 5.564 | 0.229 | 0.823 | 0.936 | 0.974 |
| Almalioglu et al. [91] | 2019 | Unsup | 80 | 0.150 | 1.141 | 5.448 | 0.216 | 0.808 | 0.939 | 0.975 |

scenes from stereo image pairs. Besides, a disparity map instead of the depth map is predicted by the designed network, and the ground truth disparity is used for supervision:

$$\mathcal{L}(I, I^*) = \frac{1}{N}\sum_{i}^{N} \|Dis_i - Dis_i^*\|_1, \tag{5}$$

where $Dis_i$ stands for the predicted disparity of pixel $i$, and $Dis_i^*$ is the corresponding ground truth. Considering the slow convergence and local optimal solutions caused by minimizing mean squared error in log-space during training, Fu et al. [64] regarded the monocular depth estimation as an ordinal regression problem. As the uncertainty of the predicted depth values increase along with the ground truth depth values, it is better to allow larger estimation errors when predicting larger depth values, which cannot be well solved by the uniform discretization (UD) strategy. Therefore, a spacing-increasing discretization (SID) strategy was proposed in ref. [64] to discretize depth and optimize the training process. To improve the transportability of depth network on different cameras, Facil et al. [22] introduced the camera model into the depth estimation network, improving the generalization capabilities of networks. Although the above methods achieve outstanding accuracy, a large number of parameters in these works limit the applications of their network in practice, especially on embedded systems. Hence, Wofk et al. [65] addressed this problem by designing a lightweight auto-encoder network framework. Meanwhile, the network pruning is applied to reduce computational complexity and improves real-time performance.

**Methods based on conditional random fields** Instead of using an additional network to refine the results in ref. [35], Li et al. [51] proposed a refinement method based on the hierarchical conditional random fields (CRFs), which is also widely used for semantic segmentation [95, 96]. Because of the continuous characteristics of depth between pixels, CRF can refine the depth estimation by considering the depth information of neighboring pixels, so that the CRF model is widely applied in the depth estimation [51–53]. In ref. [51], a deep CNN framework is designed to regress the depth map from multi-level image patches at the super-pixel level. Then, the depth map is refined from super-pixel level to pixel level via hierarchical CRF, and the energy function is

$$\mathbf{E(d)} = \sum_{i\in S}\phi_i(d_i) + \sum_{(i,j)\in\varepsilon_s}\phi_{ij}(d_i, d_j) + \sum_{c\in P}\phi_c(d_c), \tag{6}$$

where $S$ stands for the set of super-pixels, $\varepsilon_s$ refers to the set of super-pixel pairs that share a common boundary, and $P$ denotes the set of pixel-level patches. **E(d)** consists of three parts: (1) a data term for calculating the quadratic distance between the depth value $d$ and the network regressed depth $\bar{d}$;

(2) a smoothness term for enforcing relevance between neighboring super-pixels; (3) an auto-regression model for describing the local relevant structure in the depth map. In the same year, a similar framework exploring deep CNN with continuous CRF, called deep convolutional neural fields, was proposed by Liu et al. [52] to tackle the problem of monocular depth estimation. Meanwhile, a super-pixel pooling method was proposed by them to speed up the convolutional network, and it helps to design the deeper network to improve the accuracy of depth estimation. Wang et al. [53] present a framework jointly estimate the pixel-level depth map and semantic labels from a single image. Because of the structural consistency between the depth map and semantic labels, the interactions between the depth and semantic information are utilized to improve the performance of depth estimation. The depth and semantic prediction tasks are jointly trained by the supervised signal:

$$\mathcal{L}(I, I^*) = \frac{1}{N}\sum_{i}^{N}(\log(d_i) - \log(d_i^*))^2 - \lambda\frac{1}{N}\sum_{i}^{N}\log(P(l_i^*)), \quad (7)$$

$$P(l_i^*) = \exp(z_{i,l_i^*})/\sum_{l_i}\exp(z_{i,l_i}), \quad (8)$$

where $l_i^*$ stands for the ground truth of semantic labels. Meanwhile, $l_i$ refers to the predicted labels. $z_{i,l_i^*}$ denotes the output of the semantic node. To further refine the estimated depth, Wang et al. [53] also introduced a two-layer hierarchical CRF to update the depth details by extracting frequent templates for each semantic category, which lead to the fact that their methods cannot perform well as the number of classes increases. Therefore, Mousavian et al. [57] present a coupled framework for simultaneously estimating depth maps and semantic labels from a single image, and these two tasks share high-level feature representation of images extracted by CNN. A fully connected CRF is used and coupled with deep CNN to enhance the interactions between depth maps and semantic labels. Hence, their method is trained in an end-to-end manner and 10 times faster than that reported in ref. [53]. Zhang et al. [61] proposed a task-attentional module to encapsulate the interaction and improve the performance of networks, which is different from previous works [53,57]. Similar to ref. [57], Xu et al. [62] also integrated the continuous CRF model into deep CNN framework for end-to-end training. Besides, a structured attention model coupled with the CRF model is proposed in ref. [62] to strengthen the information transfer between corresponding features. The random forests (RFs) model is also introduced to monocular depth estimation tasks and efficiently enforces the accuracy of depth estimation [58].

**Methods based on adversarial learning** Because of the outstanding performance on data generation [97], the adversarial learning proposed in ref. [98] has become a hot research direction in recent years. Varieties of algorithms, theories, and applications have been widely developed, which was reviewed in ref. [99]. The frameworks of adversarial learning in depth estimation are shown in Figure 2. Different kinds of adversarial learning frameworks based on ref. [98], like stacked GAN [100], conditional GAN [101] and cycle GAN [102], are introduced into depth estimation tasks and have a positive impact on the depth estimation [60, 63, 75]. Jung et al. [60] introduced the adversarial learning into monocular depth estimation tasks. The generator consists of a Global Net and a Refinement Net, and these networks are designed to estimate the global and local 3D structures from a single image. Then, a discriminator is used to distinguish the predicted depth maps from the real ones, and this form is commonly used in supervised methods. The confrontation between generator $G$ and discriminator $D$ facilitates the training of the framework based on the min-max problem:

$$\min_G \max_D \mathbb{E}_{x\sim P_{gt}}[\log D(x)] + \mathbb{E}_{\hat{x}\sim P_G}[\log(1 - D(\hat{x}))], \quad (9)$$

where $x$ is the ground truth depth map, and $\hat{x}$ refers to the depth map predicted by generator. Similarly, conditional GAN was also used in ref. [63] for monocular depth estimation. The difference from ref. [60] is that a secondary

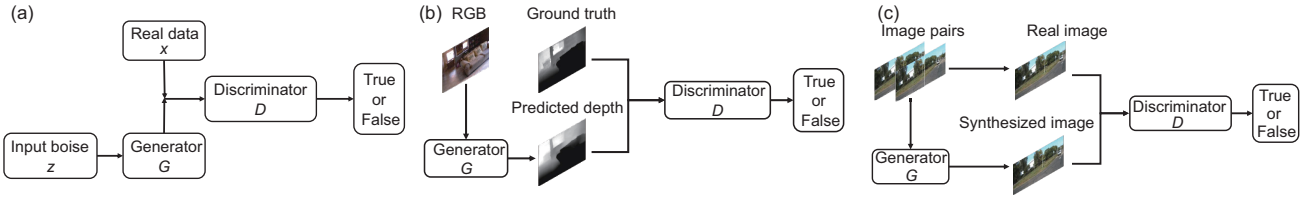


**Figure 2** (Color online) (a) The framework of raw GAN. The generator of raw GAN [98] has the ability to generate the data from a vector **z**, and the discriminator is designed to distinguish the real and fake data. Finally, the data generated by generator has the same data distribution as the real data. (b) The framework of supervised methods based on GAN. In the GAN-based supervised methods [60], the depth maps predicted by generator (depth network) and real depth maps are sent to discriminator during training. (c) The framework of unsupervised and semi-supervised methods based on GAN. In the GAN-based unsupervised and semi-supervised methods [75,83], owing to the lack of real dense depth maps, the RGB images synthesized by view reconstruction algorithm in the generator and real images instead of depth maps are sent to discriminator, and the generator takes image pairs, like the image snippets in unsupervised methods or the stereo image pairs in semi-supervised methods, to estimate the depth maps from single images and synthesize RGB images.

GAN is introduced to get a more refined depth map based on the image and coarse estimated depth map.

Because of being supervised by the ground truth, the supervised methods can effectively learn the functions to map 3D structures and their scale information from single images. However, these supervised methods are limited by the labeled training sets, which are hard and expensive to acquire.

### 3.2 Unsupervised monocular depth estimation

Instead of using the ground truth, which is expensive to acquire, the geometric constraints between frames are regarded as the supervisory signal during the training process of the unsupervised methods.

**A basic model for unsupervised methods** The unsupervised methods are trained by monocular image sequences, and the geometric constraints are built on the projection between neighboring frames:

$$p_{n-1} \sim \mathbf{K}T_{n\to n-1}D_n(p_n)\mathbf{K}^{-1}p_n, \tag{10}$$

where $p_n$ stands for the pixel on image $I_n$, and $p_{n-1}$ refers to the corresponding pixel of $p_n$ on image $I_{n-1}$. $\mathbf{K}$ is the camera intrinsics matrix, which is known. $D_n(p_n)$ denotes the depth value at pixel $p_n$, and $T_{n\to n-1}$ represents the spatial transformation between $I_n$ and $I_{n-1}$. Hence, if $D_n(p_n)$ and $T_{n\to n-1}$ are known, the correspondences between the pixels on different images ($I_n$ and $I_{n-1}$) are established by projection function. Inspired by this constraint, Zhou et al. [43] designed a depth network to predict the depth map $\hat{D}_n$ from a single image $I_n$, and a pose network to regress the transformation $\hat{T}_{n\to n-1}$ between frames ($I_n$ and $I_{n-1}$). Based on the output of networks, the pixel correspondences between $I_n$ and $I_{n-1}$ are built up:

$$p_{n-1} \sim \mathbf{K}\hat{T}_{n\to n-1}\hat{D}_n(p_n)\mathbf{K}^{-1}p_n. \tag{11}$$

Then, the photometric error between the corresponding pixels is calculated as the geometric constraints. Zhou et al. were inspired by ref. [103] to use a view synthesis as a metric, and the reconstruction loss is formulated as

$$\mathcal{L}_{vs} = \frac{1}{N}\sum_{p}^{N}|I_n(p) - \hat{I}_n(p)|, \tag{12}$$

where $p$ indexes over pixel coordinates. $\hat{I}_n(p)$ denotes the reconstructed frame. The structure similarity based on structural similarity (SSIM) is also introduced into $\mathcal{L}_{vs}$ to quantify the differences between reconstructed and target images:

$$\mathcal{L}_{vs} = \alpha\frac{1-\mathrm{SSIM}(I_n - \hat{I}_n)}{2} + (1-\alpha)|I_n - \hat{I}_n|, \tag{13}$$

where $\alpha$ is a balance factor. Besides, the recent work [104] has proven that it is more efficient to calculate the minimum value of the reconstruction error than the mean, which has been applied in refs. [86, 105]. The view reconstruction algorithm is applied to reconstruct the frame $\hat{I}_n(p)$ from $I_{n-1}$ based on the projection function, as shown in Figure 3. An edge-aware depth smoothness loss similar to refs. [27, 106] is adopted to encourage the local smooth of depth map:

$$\mathcal{L}_{\mathrm{smooth}} = \frac{1}{N}\sum_{p}^{N}|\nabla D(p)|\cdot(\mathrm{e}^{-|\nabla I(p)|})^{\mathrm{T}}, \tag{14}$$

where T refers to the transpose operation. Although the depth network is coupled with pose network during training, as shown in Figure 3, they can be used independently during testing. The above eqs. (11)-(14) form the basic framework of the unsupervised methods.

**Methods based on explainability mask** The view reconstruction algorithm based on projection function relies on the static scenario assumption, i.e., the position of dynamic objects on neighboring frames does not satisfy the projection function, which affects the photometric error and training process. Therefore, masks are widely used to reduce the influence of dynamic objects and occlusions on view reconstruction loss $\mathcal{L}_{vs}$ (eq. (12)). In refs. [43,80], a mask network

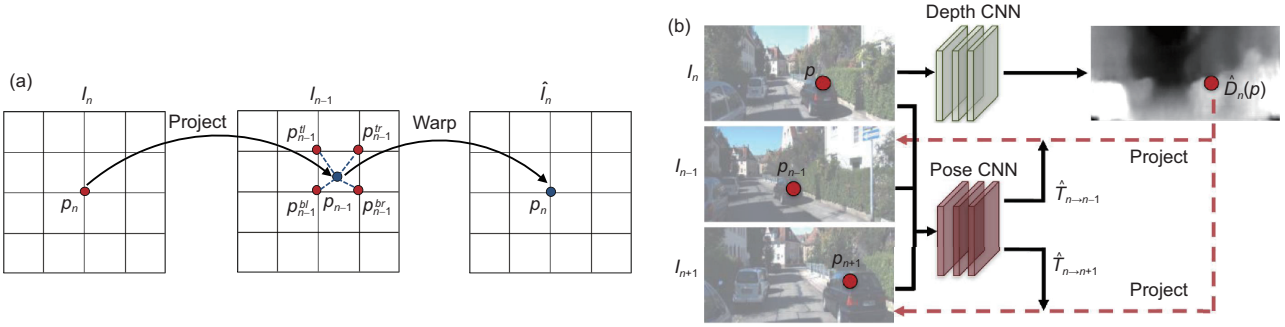


**Figure 3** (Color online) (a) An illustration of image warping process. The image warping process for view reconstruction in unsupervised methods [29,43,44]. (b) An illustration of unsupervised monocular depth. A general framework of unsupervised monocular methods proposed in ref. [43]. During training, the depth $D_n$ and pose $\hat{T}_{n\to n-1}$ predicted by the depth network and pose network are used to establish the projection relationship between $I_n$ and $I_{n-1}$, and then $\hat{I}_n$ is reconstructed by the image warping process based on projection. The differences between real $I_n$ and reconstructed $\hat{I}_n$ images are calculated to supervise the training of networks.

was designed to reduce the effects of dynamic objects and occlusions on view reconstruction through:

$$\mathcal{L}_{vs}^{M} = \frac{1}{N}\sum_{p}^{N} M|I_n(p) - \hat{I}_n(p)|, \tag{15}$$

where $M$ refers to the explainability mask predicted by a mask network. Since there is no direct supervision for $M$, training with the above loss $\mathcal{L}_{vs}^{M}$ would result in a trivial solution of the network predicting $M$ to be zero, which perfectly minimizes the loss. Therefore, a regularization term $\mathcal{L}_{\text{reg}}(M)$ is used to encourage nonzero predictions by minimizing the cross-entropy loss with constant label 1 at each pixel location. Besides, Vijayanarasimhan et al. [80] designed an object mask network to estimate the dynamic objects. The difference from ref. [43] is that the object motion is regressed together with the camera pose and used to calculate the optical flow. Based on ref. [43], Yang et al. [81] introduced a surface normal and a depth-normal consistency term for the unsupervised framework to enhance the constraints on depth estimation. The mutual conversion between depth and normal is solved by designing a depth-to-normal layer and a normal-to-depth layer in the depth network. As a result, the depth network achieves higher accuracy than ref. [43]. Mahjourian et al. [50] explored the geometric constraints between the depth map of consecutive frames. They proposed an ICP loss term to enforce consistency of the estimated depth maps, and their total network framework (including mask network, pose network and depth network) are similar to ref. [43].

Although the mask estimation based on deep neural network is widely used in previous works [43,50,80,81] and effectively reduces the effects of dynamic objects and occlusion on reconstruction errors, it not only increases the amount of computations, but also complicates network training. Therefore, in refs. [44,85], the geometry-based masks are designed to replace the masks based on deep learning and have a better effect on depth estimation. Sun et al. [84] proposed a cycle-consistent loss term to make full use of the sequence information. Wang et al. [85] carefully considered the blank regions on the reconstructed images caused by view changes and the occlusion of the pixels generated during projection. They analyze the view reconstruction process and the influence of pixel mismatch on training. Hence, two masks on the projected image and the target image, called the overlap mask and the blank mask, are proposed to solve the considered problems. Besides, a more detailed mask is designed to filter the trace mismatched pixels, and experiments prove the effectiveness of the proposed masks. The mask proposed by Bian et al. [44] is also based on geometry consistency constraint. They design a self-discovered mask based on the inconsistency between the depth maps of adjacent images. Besides, a scale consistency loss term is proposed in ref. [44], and it significantly tackles the problem of scale inconsistency between different depth maps.

**Methods based on traditional visual odometry** Instead of using the pose estimated by a pose network, the pose regressed from traditional direct visual odometry is used to assist the depth estimation in ref. [30]. The direct visual odometry takes the depth map generated by the depth network and a three-frame snippet to estimate the poses between frames by minimizing the photometric error; then, the calculated poses are sent back to the training framework. Therefore, because the depth network is supervised by more accurate poses, the accuracy of depth estimation is significantly improved.

**Methods based on multi-tasks framework** Recent approaches introduce additional networks for multi-task into the basic framework, like optical flow [29, 82], object motion [86, 87] and camera intrinsics matrix [88, 89]. Hence, the geometric relationship between different tasks is used as an additional supervision signal, which strengthens the training of the entire framework. Yin and Shi [29] proposed a jointly learning framework for depth, ego-motion and optical flow tasks. The proposed unsupervised framework consists of two parts: the rigid structure reconstructor for rigid scene reconstruction, and the non-rigid motion localizer for dynamic objects processing. A ResFlowNet is designed in the second part to learn the residual non-rigid flow. Therefore, the accuracy of all three tasks has been improved by separating rigid and non-rigid scenes and eliminating outliers through the proposed adaptive geometric consistency loss. Since the flow field of rigid regions in refs. [29, 80] is generated by the depth and pose estimation, errors produced by depth or pose estimation are propagated to the flow prediction. Therefore, Zou et al. [82] designed an additional network to estimate the optical flow. Besides, they proposed a cross-task consistency loss to constrain the consistency between the estimated flow (from network) and the generated flow (from depth and pose estimation). Ranjan et al. [87] further extended the multi-task framework, and the motion segmentation is jointly trained with other tasks (depth, pose, flow) in an unsupervised way. More tasks make the training process more complicated, so they introduce competitive collaboration to coordinate the training process and achieve outstanding performance. Similar to ref. [80], Casser et al. [86] also carefully considered the motions of dynamic objects in the scenes. An object motion network is introduced to predict the motions of individual objects, and this network takes the segmented images as input. Since the above methods are based on the prerequisites of known camera intrinsic parameters, this limits the application of the network to unknown cameras. Therefore, in refs. [88, 89], they extend the pose network to estimate the camera

intrinsic parameter and further reduce the prerequisites during training.

**Methods based on adversarial learning** The adversarial learning framework is also introduced to unsupervised monocular depth estimation. Since there are no real depth maps in the unsupervised training, it is not feasible to utilize adversarial learning like eq. (12). Therefore, instead of using the discriminator to distinguish the real and predicted depth maps, the images synthesized by view reconstruction algorithms and the real images are regarded as the input of discriminator. In refs. [83, 90, 91], the generator consists of a pose network and a depth network, and the output of networks is used to synthesize images by view reconstruction. Then, a discriminator is designed to distinguish the real and predicted depth maps. Since temporal information helps to improve the performance of the network, the LSTM module is introduced to the pose network and depth network to contact contextual information in refs. [90, 91]. Furthermore, Li et al. [90] designed an additional network to eliminate the shortcoming of view reconstruction algorithm, which is similar to ref. [43]. In order to get more 3D cues, the information between frames extracted by LSTM and the single images are adopted together to depth estimation.

Compared with supervised and semi-supervised methods, unsupervised methods learn the depth information from the geometric constraints instead of ground truth. Therefore, the training process relies on monocular sequences captured by a camera, and unsupervised learning is beneficial for the practical application of unsupervised methods. However, because of learning from monocular sequences, which do not contain the absolute scale information, unsupervised methods suffer from scale ambiguity, scale inconsistency, occlusions and other problems.

### 3.3 Semi-supervised monocular depth estimation

Since there is no need for ground truth during training, the performance of unsupervised methods is still far from the supervised methods. Besides, the unsupervised methods also suffer from various problems, like scale ambiguity and scale inconsistency. Therefore, the semi-supervised methods are proposed to get higher estimation accuracy while reducing the dependence on the expensive ground truth. Besides, the scale information can be learned from the semi-supervised signals.

Training on stereo image pairs is similar to the case of monocular videos, and the main difference is whether the transformation between two frames (left-right images or front-back images) is known. Therefore, some studies regard the framework based on stereo image pairs as unsupervised methods [23], while others treat them as semi-supervised methods [102]. In this review, we consider them the semi-supervised methods, and the poses between left-right images are the supervised signals during training.

**A basic model for semi-supervised methods** Semi-supervised methods trained on stereo image pairs estimate the disparity maps (inverse depth maps) between the left and right images. Then, the disparity map $Dis$ calculated from predicted inverse depth is used to synthesize the left image from the right image by inverse warping, as shown in Figure 4. Similar to the unsupervised methods, the differences between the synthesized images $I_w$ and real images $I_l$ are used as a supervised signal and to constrain the training process:

$$\begin{aligned}\mathcal{L}_{\text{recons}} &= \sum_p \|I_l(p) - I_w(p)\|^2 \\ &= \sum_p \|I_l(p) - I_r(p + Dis(p))\|^2,\end{aligned} \tag{16}$$

where $I_r$ is the corresponding right images. The depth map $d$ can be transferred from the predicted disparity map through: $d = fB/D$, where $f$ is the local length of cameras, and $B$ refers to the distance between left and right cameras. Based on the above framework, Garg et al. [23] also used a smoothness loss term to improve the continuities of disparity maps. Godard et al. [27] improved both the above network framework and the loss functions. The right disparity map $Dis^r$ is

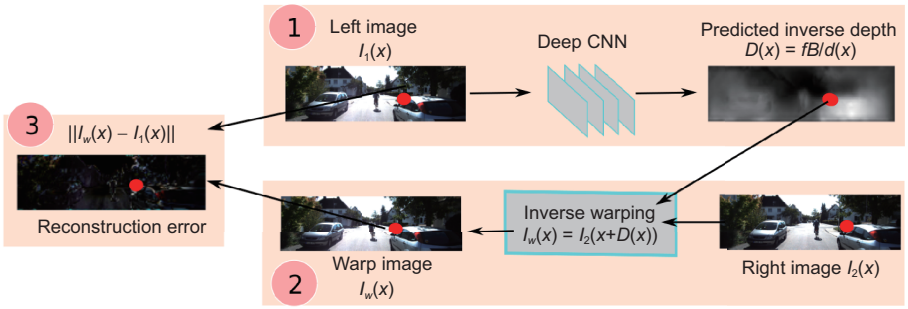


**Figure 4** (Color online) The general framework of semi-supervised monocular depth estimation based on stereo image pairs, which is proposed in ref. [23]. The depth network takes the left image to predict its pixel-level inverse depth map (or disparity map), and the predicted inverse depth map is used to reconstruct the left image from the right image by the inverse warping algorithm. The reconstruction error is calculated to supervise the training process.

predicted together with the left disparity map $Dis^l$ and used to reconstruct the right image from left image. Besides, they present a left-right disparity consistency loss to constrain the consistency between left and right disparities:

$$\mathcal{L}_{lr} = \frac{1}{N}\sum_{p} |Dis^l(p) - Dis^r(p + Dis^l(p))|. \tag{17}$$

Besides, the SSIM [107] is introduced to strengthen the structure similarity between the synthesised images and real images, and the loss function is similar to eq. (13). As a result, the experiments demonstrate the effectiveness of these improvements, and the performance outperforms the previous works [23]. Considering that above framework suffers from occlusions and left image border, a framework based on trinocular assumptions is proposed in ref. [67]. Ramirez et al. [68] proposed a framework for joint depth and semantic prediction tasks. An additional decoder stream is designed to estimate semantic labels and trained in a supervised way. Furthermore, a cross-domain discontinuity term based on the predicted semantic image is applied to improve the smoothness of the predicted depth map, which shows a better performance than the previous smoothness loss terms (like eq. (11)). Similarly, Chen et al. [74] also leveraged semantic segmentation to improve the monocular depth estimation. In ref. [74], depth estimation and semantic segmentation share the same network framework, and switch by condition. A novel left-right semantic consistency term is proposed to perform region-aware depth estimation and improves the accuracy and robustness of both tasks.

**Methods based on stereo matching** Luo et al. [70] present a view synthesis network based on Deep3D [108] to estimate the right image from the left image, which is different from above works. Moreover, a stereo matching network is designed to take the raw left and synthesised right images to regress the disparity map. During training, the view synthesis network is supervised by the raw right images to improve the construction quality, and the predicted disparity maps are used to reconstruct the left images from the estimated right images. Similar to ref. [70], Tosi et al. [73] also leveraged the stereo matching strategy to improve the performance and robustness of monocular depth estimation. Features from different viewpoints are synthesised by performing stereo matching, thereby achieving outstanding performance. Their network framework consists of three parts: a multi-scale feature extractor for high-level feature extraction, a disparity network for disparity map prediction, and a refinement network for disparity refinement. In comparison to ref. [70], the networks proposed in ref. [73] are jointly trained while those in ref. [70] are trained independently, thereby simplifying the complexity of training in ref. [73].

**Methods based on adversarial learning and knowledge distillation** Combining advanced network frameworks, like adversarial learning [26, 69, 75] and knowledge distillation [72], is becoming popular and can significantly improve the performance. The framework of knowledge distillation consists of two neural networks, a teacher network and a student network. Teacher network is more complex than the student network. The purpose of knowledge distillation is to transfer the knowledge learned by the teacher network to the student network, so that the functions learned by the large model are compressed into smaller and faster models. Pilzer et al. [72] followed this idea, and knowledge distillation is used to transfer information from the refinement network to the student network. Considering the effectiveness of training with synthetic images, Zhao et al. [78] adopted the framework of cycle GAN for the transformation between the synthetic and real domains to expand the data set, and propose a geometry-aware symmetric domain adaptation network (GASDA) to make better use of the synthetic data. Their network learns from the ground truth labels in synthetic domain as well as the epipolar geometry of the real domain, thereby achieving competitive results. Wu et al. [79] improved the architecture of the generator by utilizing a spatial correspondence module for feature matching and an attention mechanism for feature re-weighting.

**Methods based on sparse ground truth** To strengthen the supervised signals, the sparse ground truth is widely incorporated into the training framework. Kuznietsov et al. [48] adopted the ground truth depth collected by LIDAR for semi-supervised learning. Besides, both the left and right depth maps $(D_l, D_r)$ are estimated by CNNs, and the supervision signal based on LIDAR data $(G_l, G_r)$ is formulated as

$$\begin{aligned}\mathcal{L}_{\text{recons}} &= \sum_{p\in\Omega_{Z,l}} \|D_l(p) - G_l(p)\|_\delta \\ &= \sum_{p\in\Omega_{Z,r}} \|D_r(p) - G_r(p)\|_\delta,\end{aligned} \tag{18}$$

where $\Omega_{Z,l}$ refers to the set of pixels with available ground truth, and $\|*\|_\delta$ denotes the Berhu norm [93]. Similarly, based on ref. [27], He et al. [71] introduced the loss between the predicted depth maps and LIDAR data as an additional signal. Moreover, the physical information is also adopted into the semi-supervised methods. Fei et al. [31] used the global orientation computed from inertial measurements as a priori information to constrain the normal vectors to surfaces of objects. Generally, normal vectors to surfaces of objects are parallel or perpendicular to the direction of gravity, and they can easily calculate from the estimated depth map. Therefore, this physical priori significantly improves the accuracy of depth estimation.

Semi-supervised methods achieve a better accuracy than unsupervised methods because of the semi-supervised signals, and the scale information can be learned from these signals. However, the accuracy of semi-supervised methods relies heavily on the ground truth, like pose and LIDAR data, although they are easier to get than expensive dense depth maps.

### 3.4 Applications

The monocular depth estimation based on deep learning has been widely applied in SLAM (or VO) to improve the mapping process, recover the absolute scale, and replace the RGB-D sensor in dense mapping. (1) Improving the map. Loo et al. [109] introduced the monocular depth prediction into the SVO framework [110], and the depth value predicted by deep neural networks is used to initialize the mean and variance of the depth at a feature location. Therefore, the depth uncertainty during mapping is effectively reduced with the help of introducing depth prediction, thereby improving the map built by CNN-SVO. (2) Scale recovery. Since the depth neural network can predict the depth containing absolute scale information from a single image, the scale ambiguity and scale drift of monocular VO methods [111] can be effectively solved with the help of deep learning-based depth estimation. Yin et al. [112] and Yang et al. [113] followed this idea and leveraged the depth estimation based on deep learning to recover the absolute scale of monocular VO. (3) Replace the RGB-D sensor. As reviewed by ref. [21], most of dense SLAM methods take RGB-D sensor to build the dense maps of scenes. Compared with RGB-D sensors, depth networks can generate the accurate and dense depth maps from the single images captured by monocular cameras. Besides, monocular cameras have the advantages of small size, low power consumption, and easy access. Therefore, Tateno et al. [13] proposed a method that introduce the deep depth estimation into dense monocular reconstruction, and their methods also demonstrate the effectiveness of deep learning-based depth prediction in the absolute scale recovery.

## 4 Discussion

In general, we think that the development of monocular depth estimation will still focus on improving the accuracy, transferability, and real-time performance.

**Accuracy** Most of the previous works mainly focus on improving the accuracy of depth estimation by adopting new loss functions or network frameworks, as shown in Table 1. Several well-known network frameworks, like LSTM, VAE, GANs, have shown their effectiveness in improving the performance of depth estimation. Therefore, with the development of deep neural networks, trying new network frameworks, like 3D convolution [114], graph convolution [115], attentional mechanism [116] and knowledge distillation [117], may get satisfactory results. Although the unsupervised methods do not rely on ground truth during training, their accuracy is far from the current most effective semi-supervised methods, as shown in Table 2. Finding a more efficient geometric constraint to improve the unsupervised methods [104] may be a good direction. For example, the target-level dynamic object motion estimation combined with geometry-based mask, will be an effective solution to the impact of dynamic objects and occlusions on view reconstruction. Besides, the unsupervised methods training on monocular videos suffer from scale ambiguity and scale inconsistency. Although some loss terms are proposed to constrain the scale consistency, this problem is not solved well. Since the semantic information is mainly used to constrain the smoothness of depth map during training, it will be a good research direction for solving monocular scale ambiguity by learning the scale from semantic information. Moreover, the multi-task joint training combining with the geometric relationship between tasks is also a proven method that is worthy of further study. To get a state-of-the-art result, the network framework is becoming more and more complicated, and the loss terms are becoming more complicated, which make the training of the network difficult. Furthermore, the increase of loss terms will also pose a challenge on the selection of hyperparameters. A more effective way for designing deep learning-based hyperparameter setting methods is also a huge challenge. For example, estimating the intrinsics matrix of the monocular camera and the parameters of stereo cameras based on deep learning may be a promising direction.

**Transferability** Transferability refers to the performance of the same network on different cameras, different scenarios, and different datasets. The transferability of depth networks is raising increasing attention. Most of the current methods are trained and tested on the same dataset, thereby achieving a satisfactory result. However, the training set and testing set in different domains or collected by different cameras often lead to severe performance degradation. Incorporating camera parameters into depth estimation framework and leveraging domain adaptation technology during training will significantly improve the transferability of depth network, and they are becoming a hot topic recently.

**Real-time performance** Although deeper networks show outstanding performance, they require more computation time to complete estimation tasks, which is a great challenge for their applications. The ability of depth estimation networks to run in real-time on embedded devices will have significant implications for their practical applications. There-

fore, the development of lightweight networks based on supervised, semi-supervised and unsupervised learning will be a promising direction, and there are not much related researches in this field at present. As the number of parameters of the lightweight network is smaller, this affects the performance of the network. Therefore, it is a worthwhile subject to improve accuracy while ensuring real-time performance.

In addition, there is very little researches on the mechanism of monocular depth estimation methods based on deep learning, like what depth networks have learned and what depth cues they exploit. For example, the research in ref. [47] focuses on the cues of neural network learning depth from a single image, and its experiments have shown that current depth networks ignore the apparent size of known obstacles, which is different from how humans perceive depth information. Therefore, studying the mechanism of depth estimation is a promising direction, which may effectively improve the accuracy, transferability and real-time performance. The application of monocular depth estimation in environmental perception [21] and control [118, 119] of autonomous robots is also a direction worthy of research.

## 5 Conclusion

In this review, we aim to contribute to this growing area of research in deep learning-based monocular depth estimation. Therefore, we survey the related works of monocular depth estimation from the aspect of training manner, including supervised, unsupervised as well as semi-supervised learning, combining with the application of loss functions and network frameworks. In the end, we also discuss the current hot topics as well as challenges and provide some valuable ideas and promising directions for future researches.

*This work was supported by the National Key Research and Development Program of China (Grant No. 2018YFC0809302), the National Natural Science Foundation of China (Grant Nos. 61988101, 61751305 and 61673176), the Fundamental Research Funds for the Central Universities (Grant No. JKH012016011), and the Programme of Introducing Talents of Discipline to Universities (the "111" Project) (Grant No. B17017).*

1 Hu G, Huang S, Zhao L, et al. A robust RGB-D SLAM algorithm. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura: IEEE, 2012. 1714–1719
2 Zhu Z S, Su A, Liu H B, et al. Vision navigation for aircrafts based on 3D reconstruction from real-time image sequences. Sci China Tech Sci, 2015, 58: 1196–1208
3 Chai X, Gao F, Qi C K, et al. Obstacle avoidance for a hexapod robot in unknown environment. Sci China Tech Sci, 2017, 60: 818–831
4 Park S J, Hong K S, Lee S. RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, 2017. 4980–4989
5 Ullman S. The interpretation of structure from motion. Proc R Soc Lond B, 1979, 203: 405–426
6 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. Remote Sens, 2013, 5: 6880–6898
7 Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Trans Robot, 2015, 31: 1147–1163
8 Szeliski R, Kang S R. Shape ambiguities in structure from motion. IEEE Trans Pattern Anal Machine Intell, 1997, 19: 506–512
9 Zou L, Li Y. A method of stereo vision matching based on OpenCV. In: 2010 International Conference on Audio, Language and Image Processing. Shanghai: IEEE, 2010. 185–190
10 Cao Z L, Yan Z H, Wang H. Summary of binocular stereo vision matching technology (in Chinese). J Chongqing Univ Tech (Nat Sci), 2015, 29: 70–75
11 Benosman R, Manière T, Devars J. Multidirectional stereovision sensor, calibration and scenes reconstruction. In: Proceedings of 13th International Conference on Pattern Recognition. Vienna: IEEE, 1996. 161–165
12 Ramírez-Hernández L R, Rodríguez-Quiñonez J C, Castro-Toscano M J, et al. Improve three-dimensional point localization accuracy in stereo vision systems using a novel camera calibration method. Int J Adv Robot Syst, 2020, 17: 172988141989671
13 Tateno K, Tombari F, Laina I, et al. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017. 6243–6252
14 Yoneda K, Tehrani H, Ogawa T, et al. Lidar scan feature for localization with highly precise 3D map. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. Dearborn: IEEE, 2014. 1345–1350
15 Zhang F, Zhu X, Ye M. Fast human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 3517–3526
16 Pang J, Chen K, Shi J, et al. Libra R-CNN: Towards balanced learning for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 821–830
17 Lyu H, Fu H, Hu X, et al. ESNet: Edge-based segmentation network for real-time semantic segmentation in traffic scenes. In: 2019 IEEE International Conference on Image Processing (ICIP). Taipei: IEEE, 2019. 1855–1859
18 Zhao Z Q, Zheng P, Xu S T, et al. Object detection with deep learning: A review. IEEE Trans Neural Netw Learning Syst, 2019, 30: 3212–3232
19 Ghosh S, Das N, Das I, et al. Understanding deep learning techniques for image segmentation. ACM Comput Surv, 2019, 52: 1–35
20 Rawat W, Wang Z. Deep convolutional neural networks for image classification: A comprehensive review. Neur Computat, 2017, 29: 2352–2449
21 Tang Y, Zhao C, Wang J, et al. An overview of perception and decision-making in autonomous systems in the era of learning. 2020, arXiv: 2001.02319
22 Facil J M, Ummenhofer B, Zhou H, et al. CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 11826–11835
23 Garg R, Vijay Kumar B G, Carneiro G, et al. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In: Leibe B, Matas J, Sebe N, et al., eds. Computer Vision-ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9912. Cham: Springer, 2016. 740–756
24 Wang R, Pizer S M, Frahm J M. Recurrent neural network for (un-)supervised learning of monocular video visual odometry and depth. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 5555–5564

25 Chakravarty P, Narayanan P, Roussel T. GEN-SLAM: Generative modeling for monocular simultaneous localization and mapping. In: 2019 International Conference on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 147–153

26 Aleotti F, Tosi F, Poggi M, et al. Generative adversarial networks for unsupervised monocular depth prediction. In: Leal-Taixé L, Roth S, eds. Computer Vision-ECCV 2018 Workshops. ECCV 2018. Lecture Notes in Computer Science, vol 11129. Cham: Springer, 2018. 337–354

27 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. Honolulu, 2017. 270–279

28 Zhan H, Garg R, Saroj Weerasekera C, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018. 340–349

29 Yin Z, Shi J. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018. 1983–1992

30 Wang C, Miguel Buenaposada J, Zhu R, et al. Learning depth from monocular videos using direct methods. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018. 2022–2030

31 Fei X, Wong A, Soatto S. Geo-supervised visual depth prediction. IEEE Robot Autom Lett, 2019, 4: 1661–1668

32 Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3354–3361

33 Mayer N, Ilg E, Hausser P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016. 4040–4048

34 Zhao C, Tang Y, Sun Q. Deep direct visual odometry. 2019, arXiv:1912.05101

35 Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems. 2014. 2366–2374

36 Chen X, Ma H, Wan J, et al. Multi-view 3D object detection network for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017. 1907–1915

37 Wang P, Chen P, Yuan Y, et al. Understanding convolution for semantic segmentation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe: IEEE, 2018. 1451–1460

38 Chang M F, Lambert J, Sangkloy P, et al. Argoverse: 3D tracking and forecasting with rich maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 8748–8757

39 Xue F, Wang X, Li S, et al. Beyond tracking: Selecting memory and refining poses for deep visual odometry. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 8575–8583

40 Clark R, Wang S, Wen H, et al. VINet: Visual-inertial odometry as a sequence-to-sequence learning problem. In: Thirty-First AAAI Conference on Artificial Intelligence, 2017

41 Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from RGBD images. In: Fitzgibbon A, Lazebnik S, Perona P, et al., eds. Computer Vision-ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol 7576. Berlin, Heidelberg: Springer, 2012. 746–760

42 Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Las Vegas, 2016. 3213–3223

43 Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017. 1851–1858

44 Bian J, Li Z, Wang N, et al. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In: Advances in Neural Information Processing Systems, 2019. 35–45

45 Saxena A, Min Sun A, Ng A Y. Make3D: Learning 3D scene structure from a single still image. IEEE Trans Pattern Anal Mach Intell, 2009, 31: 824–840

46 Hoiem D, Efros A A, Hebert M. Automatic photo pop-up. ACM Trans Graph, 2005, 24: 577–584

47 van Dijk T, de Croon G. How do neural networks see depth in single images? In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019. 2183–2191

48 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. Honolulu, 2017. 6647–6655

49 Kendall A, Martirosyan H, Dasgupta S, et al. End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, 2017. 66–75

50 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. Salt Lake City, 2018. 5667–5675

51 Li B, Shen C, Dai Y, et al. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015. 1119–1127

52 Liu F, Shen C, Lin G, et al. Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans Pattern Anal Mach Intell, 2016, 38: 2024–2039

53 Wang P, Shen X, Lin Z, et al. Towards unified depth and semantic prediction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015. 2800–2809

54 Shelhamer E, Barron J T, Darrell T. Scene intrinsics and depth from a single image. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. Santiago, 2015. 37–44

55 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. Santiago, 2015. 2650–2658

56 Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015. 1–9

57 Mousavian A, Pirsiavash H, Košecká J. Joint semantic segmentation and depth estimation with deep convolutional networks. In: 2016 Fourth International Conference on 3D Vision (3DV). Stanford: IEEE, 2016. 611–619

58 Roy A, Todorovic S. Monocular depth estimation using neural regression forest. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016. 5506–5514

59 Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth International Conference on 3D Vision (3DV). Stanford: IEEE, 2016. 239–248

60 Jung H, Kim Y, Min D, et al. Depth prediction from a single image with conditional adversarial networks. In: 2017 IEEE International Conference on Image Processing (ICIP). Beijing: IEEE, 2017. 1717–1721

61 Zhang Z, Cui Z, Xu C, et al. Joint task-recursive learning for semantic segmentation and depth estimation. In: Ferrari V, Hebert M, Sminchisescu C, et al., eds. Computer Vision-ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11214. Cham: Springer, 2018. 238–255

62 Xu D, Wang W, Tang H, et al. Structured attention guided convolutional neural fields for monocular depth estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018. 3917–3925

63 Lore K G, Reddy K, Giering M, et al. Generative adversarial networks for depth map estimation from RGB video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, 2018. 1177–1185

64 Fu H, Gong M, Wang C, et al. Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018. 2002–2011

65 Wofk D, Ma F, Yang T J, et al. Fastdepth: Fast monocular depth estimation on embedded systems. In: 2019 International Conference on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 6101–6108

66 Chen W, Fu Z, Yang D, et al. Single-image depth perception in the wild. In: Advances in Neural Information Processing Systems. 2016, pp. 730–738

67 Poggi M, Tosi F, Mattoccia S. Learning monocular depth estimation with unsupervised trinocular assumptions. In: 2018 International Conference on 3D Vision (3DV). Verona: IEEE, 2018. 324–333

68 Ramirez P Z, Poggi M, Tosi F, et al. Geometry meets semantics for semi-supervised monocular depth estimation. In: Jawahar C, Li H, Mori G, Schindler K, eds. Computer Vision-ACCV 2018. ACCV 2018. Lecture Notes in Computer Science, vol 11363. Cham: Springer, 2019. 298–313

69 Pilzer A, Xu D, Puscas M, et al. Unsupervised adversarial depth estimation using cycled generative networks. In: 2018 International Conference on 3D Vision (3DV). Verona: IEEE, 2018. 587–595

70 Luo Y, Ren J, Lin M, et al. Single view stereo matching. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018. 155–163

71 He L, Chen C, Zhang T, et al. Wearable depth camera: Monocular depth estimation via sparse optimization under weak supervision. IEEE Access, 2018, 6: 41337–41345

72 Pilzer A, Lathuiliere S, Sebe N, et al. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 9768–9777

73 Tosi F, Aleotti F, Poggi M, et al. Learning monocular depth estimation infusing traditional stereo knowledge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 9799–9809

74 Chen P Y, Liu A H, Liu Y C, et al. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 2624–2632

75 Feng T, Gu D. SGANVO: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks. IEEE Robot Autom Lett, 2019, 4: 4431–4437

76 Li R, Wang S, Long Z, et al. UnDeepVO: Monocular visual odometry through unsupervised deep learning. In: 2018 IEEE international conference on robotics and automation (ICRA). Brisbane: IEEE, 2018. 7286–7291

77 Wang Y, Wang P, Yang Z, et al. UnOS: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 8071–8081

78 Zhao S, Fu H, Gong M, et al. Geometry-aware symmetric domain adaptation for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 9788–9798

79 Wu Z, Wu X, Zhang X, et al. Spatial correspondence with generative adversarial network: Learning depth from monocular videos. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019. 7494–7504

80 Vijayanarasimhan S, Ricco S, Schmid C, et al. SfM-Net: Learning of structure and motion from video. 2017, arXiv: 1704.07804

81 Yang Z, Wang P, Xu W, et al. Unsupervised learning of geometry with edge-aware depth-normal consistency. 2017, arXiv: 1711.03665

82 Zou Y, Luo Z, Huang J B. Df-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In: Ferrari V, Hebert M, Sminchisescu C, et al., eds. Computer Vision-ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11209. Cham: Springer, 2018. 36–53

83 Kumar A C S, Bhandarkar S M, Prasad M. Monocular depth prediction using generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, 2018. 300–308

84 Sun Q, Tang Y, Zhao C. Cycle-SfM: Joint self-supervised learning of depth and camera motion from monocular image sequences. Chaos, 2019, 29: 123102

85 Wang G, Wang H, Liu Y, et al. Unsupervised learning of monocular depth and ego-motion using multiple masks. In: 2019 International Conference on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 4724–4730

86 Casser V, Pirk S, Mahjourian R, et al. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: Proceedings of Thirty-Third AAAI Conference on Artificial Intelligence. Honolulu, 2019. 8001–8008

87 Ranjan A, Jampani V, Balles L, et al. 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. Long Beach, 2019. 12240–12249

88 Chen Y, Schmid C, Sminchisescu C. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019. 7063–7072

89 Gordon A, Li H, Jonschkowski R, et al. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019. 8977–8986

90 Li S, Xue F, Wang X, et al. Sequential adversarial learning for self-supervised deep visual odometry. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019. 2851–2860

91 Almalioglu Y, Saputra M R U, de Gusmao P P, et al. GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In: 2019 International Conference on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 5474–5480

92 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016. 770–778

93 Zwald L, Lambert-Lacroix S. The BerHu penalty and the grouped effect. 2012, arXiv: 1207.6868

94 Mancini M, Costante G, Valigi P, et al. Fast robust monocular depth estimation for obstacle detection with fully convolutional networks. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon: IEEE, 2016. 4296–4303

95 Huang Q, Han M, Wu B, et al. A hierarchical conditional random field model for labeling and segmenting images of street scenes. In: CVPR 2011. Providence: IEEE, 2011. 1953–1960

96 Ladickỳ L, Russell C, Kohli P, et al. Associative hierarchical CRFs for object class image segmentation. In: 2009 IEEE 12th International Conference on Computer Vision. Kyoto: IEEE, 2009. 739–746

97 Zhang H, Xu T, Li H, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, 2017. 5907–5915

98 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. 2014. 2672–2680

99 Hong Y, Hwang U, Yoo J, et al. How generative adversarial networks and their variants work. ACM Comput Surv, 2019, 52: 1–43

100 Huang X, Li Y, Poursaeed O, et al. Stacked generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017. 5077–5086

101 Mirza M, Osindero S. Conditional generative adversarial nets. 2014, arXiv: 1411.1784

102 Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, 2017. 2223–2232

103 Szeliski R, Prediction error as a quality metric for motion and stereo. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. Kerkyra: IEEE, 1999. 781–788

104 Godard C, Mac Aodha O, Firman M, et al. Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019. 3828–3838

105 Bozorgtabar B, Rad M S, Mahapatra D, et al. SynDeMo: Synergistic deep feature alignment for joint learning of depth and ego-motion. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019. 4210–4219

106 Heise P, Klose S, Jensen B, et al. PM-Huber: PatchMatch with Huber regularization for stereo matching. In: Proceedings of the IEEE International Conference on Computer Vision. Sydney, 2013. 2360–2367

107 Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process, 2004, 13: 600–612

108 Xie J, Girshick R, Farhadi A. Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In: Leibe B, Matas J, Sebe N, et al., eds. Computer Vision-ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9908. Cham: Springer, 2016. 842–857

109 Loo S Y, Amiri A J, Mashohor S, et al. CNN-SVO: Improving the mapping in semi-direct visual odometry using single-image depth prediction. In: 2019 International Conference on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 5218–5223

110 Forster C, Pizzoli M, Scaramuzza D. SVO: Fast semi-direct monocular visual odometry. In: 2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong: IEEE, 2014. 15–22

111 Engel J, Koltun V, Cremers D. Direct sparse odometry. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 611–625

112 Yin X, Wang X, Du X, et al. Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields. In Proceedings of the IEEE International Conference on Computer Vision. Venice, 2017. 5870–5878

113 Yang N, Wang R, Stuckler J, et al. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 817–833

114 Cheng X, Wang P, Yang R. Learning depth with convolutional spatial propagation network. 2018, arXiv: 1810.02695

115 Li Q, Han Z, Wu X M. Deeper insights into graph convolutional networks for semi-supervised learning. In: Thirty-Second AAAI Conference on Artificial Intelligence. 2018

116 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems, 2017. 5998–6008

117 Ahn S, Hu S X, Damianou A, et al. Variational information distillation for knowledge transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. 9163–9171

118 Wu X, Tang Y, Zhang W. Stability analysis of stochastic delayed systems with an application to multi-agent systems. IEEE Trans Automat Contr, 2016, 61: 4143–4149

119 Tang Y, Wu X, Shi P, et al. Input-to-state stability for nonlinear systems with stochastic impulses. Automatica, 2020, 113: 108766