Neurocomputing 438 (2021) 14–33

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Deep learning for monocular depth estimation: A review ☆

Yue Ming [a,1], Xuyang Meng [a,1], Chunxiao Fan [a], Hui Yu [b,*]

[a] Beijing Key Laboratory of Work Safety and Intelligent Monitoring, School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, PR China
[b] School of Creative Technologies, University of Portsmouth, UK

ARTICLE INFO

*Article history:*
Received 20 October 2020
Revised 19 December 2020
Accepted 19 December 2020
Available online 5 January 2021
Communicated by Zidong Wang

*Keywords:*
Monocular depth estimation
Deep learning
Supervised learning
Unsupervised learning
Multi-task learning

ABSTRACT

Depth estimation is a classic task in computer vision, which is of great significance for many applications such as augmented reality, target tracking and autonomous driving. Traditional monocular depth estimation methods are based on depth cues for depth prediction with strict requirements, e.g. shape-from-focus/ defocus methods require low depth of field on the scenes and images. Recently, a large body of deep learning methods have been proposed and has shown great promise in handling the traditional ill-posed problem. This paper aims to review the state-of-the-art development in deep learning-based monocular depth estimation. We give an overview of published papers between 2014 and 2020 in terms of training manners and task types. We firstly summarize the deep learning models for monocular depth estimation. Secondly, we categorize various deep learning-based methods in monocular depth estimation. Thirdly, we introduce the publicly available dataset and the evaluation metrics. And we also analysis the properties of these methods and compare their performance. Finally, we highlight the challenges in order to inform the future research directions.



## 1. Introduction

Scene depth estimation plays an important role in computer vision, which enhances the perception and understanding of real three-dimensional scenes leading to a wide range of applications such as robotic navigation, autonomous driving, and virtual reality [1,53,139,145,166]. Active depth estimation methods usually utilize lasers, structured light and other reflections on the object surface to obtain depth point clouds, complete surface modeling and estimate scene depth maps [61,182]. However, obtaining dense and accurate depth maps usually requires extremely heavy costs of manpower and computing resources [101,178]. Therefore, image-based depth estimation has become the mainstream of research, and can be applied in a wide range of applications [89,135].

The evolution of image-based depth estimation is shown in Fig. 1. In the early period, researchers estimated depth maps depending on depth cues, such as vanishing points [142], focus and defocus [138], and shadow [181]. However, most of these methods were applied in constraint scenes [138,142,181]. With the development of computer vision, many hand-made features and probabilistic graph models have been proposed, such as scale-invariant feature transform (SIFT) [88], speeded up robust features (SURF) [7], pyramid histogram of oriented gradient (PHOG) [9], Conditional Random Field (CRF) [66], and Markov Random Field (MRF) [25], which were adopted to predict monocular depth maps with parameter and non-parameter learning in the machine learning process [25,66,81]. The advent of deep learning technologies has brought great advantages to image processing [47,68,148,172] especially depth estimation.

Traditional depth estimation methods of image-based depth estimation are usually based on binocular camera, which calculates the disparity of two 2D images (taken by a binocular camera) through stereo matching and triangulation to obtain a depth map [40,82,117,170,180]. However, the binocular depth estimation method requires at least two fixed cameras [185], and it is difficult to capture enough features in the image to match when the scene has less or no texture [84]. Therefore, researchers turn their attention to monocular depth estimation. Monocular depth estimation uses only one camera to obtain an image or video sequence, which does not require additional complicated equipments and professional techniques. It has vast application demands due to the availability of only one single camera in most application scenarios. Thus,there is an increasing demand for monocular depth estimation in recent years. Since monocular images lack a reliable stereo-

☆ The work presented in this paper was partly supported by Natural Science Foundation of China (Grant No. 62076030), Beijing Natural Science Foundation of China (Grant No. L201023, and No. L182033) and the Fundamental Research Funds for the Central Universities (2019PTB-001).

* Corresponding author.
*E-mail address:* hui.yu@port.ac.uk (H. Yu).
[1] Co-first author.

https://doi.org/10.1016/j.neucom.2020.12.089
0925-2312/

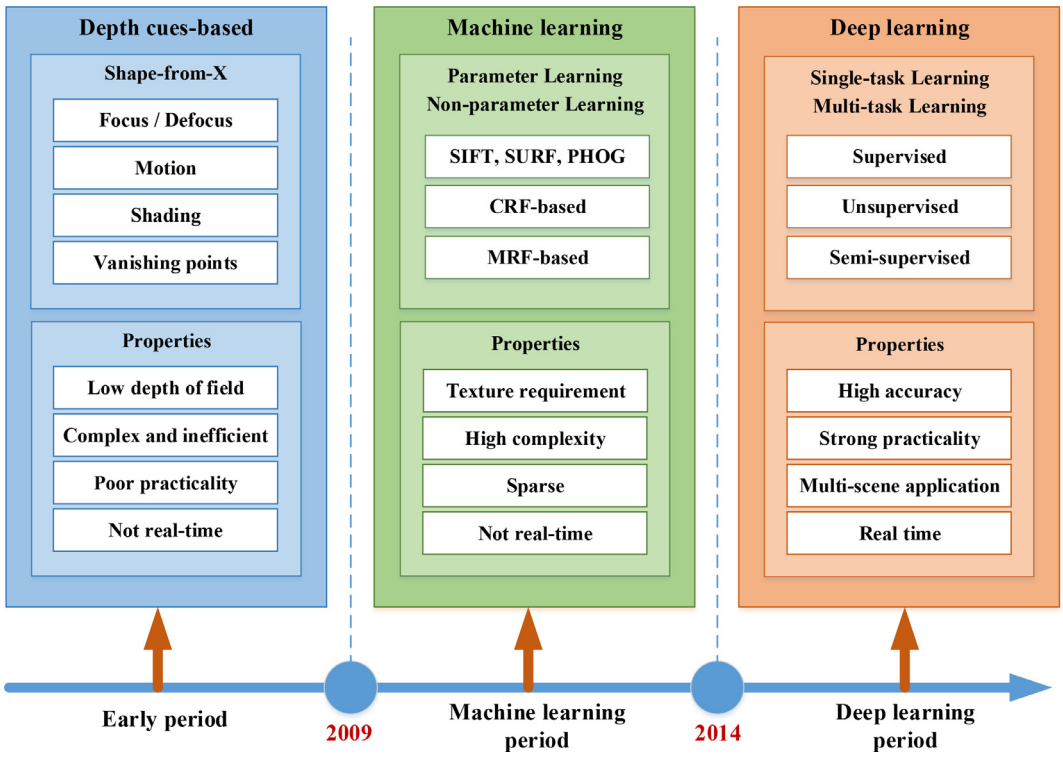



**Fig. 1.** The evolution of depth estimation. This paper divides the development of depth estimation into three periods: the early period, the machine learning period, and the deep learning period, where the depth estimation method of monocular image based on deep learning is mainly surveyed and summarized.

scopic visual relationship, it is essentially an ill-posed problem to regress depth in 3D space [102]. Therefore, researchers propose various methods for monocular depth estimation [8,67].

Monocular images adopt a two-dimensional form to reflect the three-dimensional world. However, one dimension of the scene, namely depth, has missed in the imaging process, which makes it impossible to judge the size and distance of the object, nor to judge whether the object is occluded by another object. Therefore, we need to recover the depth of the monocular image. Based on the depth map, we can judge the size and distance of the object to meet the needs of scene understanding. When the estimated depth map can reflect the three-dimensional structure of the scene, we can consider that the depth estimation method is effectiveness.

This paper focuses on the research of monocular depth estimation, which surveys deep learning-based methods in recent years, details their remarks, and compares their performances. Furthermore, this paper describes the limitations of these existing methods and briefly introduces the future trends. The remainder of this paper is as follows: Section 2 introduces some deep learning models for monocular depth estimation; Section 3 summarizes deep learning-based methods of monocular depth estimation, from training manners and task types; Section 4 introduces the common datasets and evaluation metrics of depth estimation, and then analysis their properties and compares their performance; Section 5 discusses the challenges and trends of monocular depth estimation; Conclusions are drawn in Section 6.

## 2. Deep Learning models for monocular depth estimation

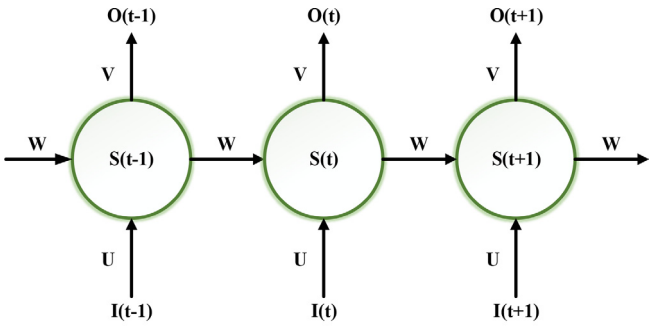
This section mainly introduces common deep learning models for monocular depth estimation: Convolutional Neural Network (CNN) [63], Recurrent Neural Network (RNN) [122], and Generative Adversarial Network (GAN) [39].

### 2.1. CNN

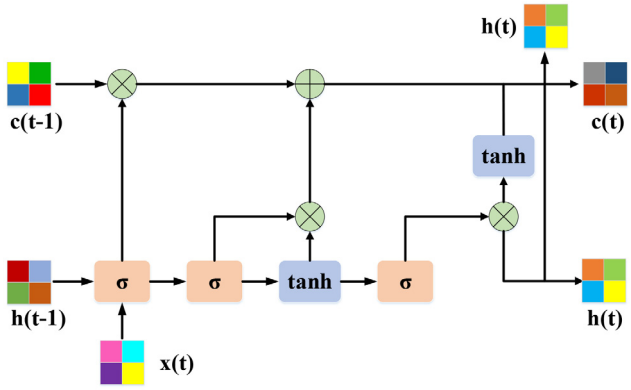
CNN can automatically extract spatial features representing depth in a scene. It is a type of feed-forward neural network, which extracts depth features and reconstructs depth maps at the same time with fewer parameters compared to traditional methods [165,159,86]. CNN mainly includes convolutional layer, pooling layer, fully connected layer and activation function, which enable CNN to learn the two-dimensional spatial features of the input image. The convolutional layer transforms the input into depth features; the pooling layer reduces the size of the input feature map in max-pooling or average-pooling manner; the fully connected layer is usually located at the end of the CNN to output the results; and the activation function is generally a continuously differentiable nonlinear function to avoid pure linear combinations. Representative CNNs include AlexNet [63], VGG [131], GoogLeNet [137], ResNet [48], DenseNet [52], and some lightweight network, such as MobileNet [51], ShuffleNet [183], and GhostNet [46], each of which is used as the backbone of the existing CNN-based depth estimation network.

### 2.2. RNN

RNN is a sequence-to-sequence model with memory capabilities [13,41] as shown in Fig. 2(a), which is introduced into monocular depth estimation so as to learn temporal features from video sequences. RNN includes three parts: input unit, hidden unit, and output unit, where the input of the hidden unit consists of the outputs of both current input unit and previous hidden unit. Furthermore, Hochreiter et al. [50] proposed a Long Short-Term Memory (LSTM) unit as shown in Fig. 2(b), which could learn long-term dependences with a three-gate structure: input gate layer, forget gate layer, and output gate layer. Representative RNNs including BiRNN [126], GRU [22], ConvLSTM [162], $G^2$-LSTM [78], ON-LSTM




(a) RNN




(b) LSTM

**Fig. 2.** (a) The basic structure of RNN, where $S$ is the internal status and the memory of the cell, $I$ is the input, $O$ is the output, and $(U, V, W)$ is the sharing parameters of the cell. (b) The basic structure of LSTM [50].

[127], Mogrifier LSTM [96] and others are introduced into deep learning models for monocular depth estimation, which are usually combined with CNNs to extract spatial–temporal features to recover depth [54,149].

### 2.3. GAN

The supervised depth estimation model needs to learn the 3D mapping and scale information from the ground truth (GT) depth maps. However, it is difficult to obtain GT depth maps in real scenes so that researchers introduced GAN [39] to generate clearer and more realistic depth maps compared to other models [177]. GAN includes two modules: the generator predicts the depth map as a depth estimation network, and the discriminator determines whether the input depth map is true or false, as shown in Fig. 3. Representative GANs are introduced into depth estimation, including conditional GAN [99], DCGAN [111], WGAN [4], stacked GAN [177], SimGAN [128], and Cycle GAN [196]. Depth estimation models with GANs can provide generation adversarial constraints for the estimated depth maps and the GT depth maps [32,45,58].

## 3. Deep learning methods for monocular depth estimation

Deep neural networks have played an important role in various areas with their powerful feature learning ability. Monocular depth estimation based deep learning is a task of learning depth maps from a single 2D color image through a deep neural network, which was firstly proposed by Eigen et al. [29] in 2014. It was a coarse-to-fine framework, where the coarse network learned the global depth on the entire image to obtain a rough depth map and the fine network learned the local features to refine the depth map, as shown in Fig. 4. Since then, many researchers have carried out deep learning methods for monocular depth estimation [28,30,36,69,83,169,174,189].

The framework of monocular depth estimation based on deep learning is an encoder-decoder network, with the RGB image input and depth map output, as shown in Fig. 5. The encoder network consists of convolution and pooling layers to capture the depth features, and the decoder network includes deconvolution layers to regress the estimated pixel-level depth map, with the same size as the input. Additionally, in order to preserve the features of each scale, the corresponding layers of encoder and decoder are concatenated with skip-connections. The entire network is constrained and trained by the depth loss functions and converges when the desired depth map is generated.

Deep learning methods for monocular depth estimation often utilize gradient descent to train deep neural networks, and obtain a local minimum finally. The best local minimum depends on initialization and specific parameter settings. In the initialization process, it is generally necessary to resize the image to meet the needs of network learning. In addition, it also need to set the initial learning rate, optimizer parameters, batchsize and mini-batchsize, to learn and save image features. The commonly used learning method is stochastic gradient descent, and the optimizer is Adam. When the gradient no longer changes and the loss function becomes stable, the network converges.

Compared with traditional methods, deep learning methods for monocular depth estimation construct the multi-layer neural network to learn deep features, which has higher accuracy. When there is small occlusion in the monocular image or part of the ground-truth depth is missing, the deep learning methods can still estimate the depth of the scene, and have low errors; when there is large occlusion in presence in the scene or there is no ground-truth depth, deep learning methods can learn the depth of the scene by adding network constraints. In short, deep learning methods for monocular depth estimation have shown strong robustness.

This section reviews and summarizes deep learning methods for monocular depth estimation from 2014 to 2020, which was classified into two different perspectives: the training manners with supervised, unsupervised and semi-supervised manner, and the tasks with single-task and multi-task learning of depth estimation models. The overall diagram of monocular depth estimation based on deep learning is drawn in Fig. 6.

### 3.1. Training manners

The supervised monocular depth estimation network estimates the depth maps by learning the scene structure information from the GT depth maps. The cost of obtaining the GT depth maps is very high, so that some monocular depth estimation networks need to be trained with less or no GT to reconstruct depth maps, which are the semi-supervised or unsupervised learning methods. This section will review and classify deep learning methods from the perspective of training manners: supervised, unsupervised, and semi-supervised models for monocular depth estimation.

#### 3.1.1. Supervised learning methods

Supervised learning networks for monocular depth estimation are trained with the GT depth maps as shown in Fig. 7. The purpose of learning is to penalize the errors between the predictions and GT depth maps constrained by the loss functions formulated in Table 1, where the $log(d)$ as Eq. (1) is based on $log$ depth [28], and the reverse Huber (Berhu) function as Eq. (1) combines the $L_1$ and $L_2$ norms at the same time to reduce the influence of error changes on the range of weights proposed by Laina et al. [69]. That is, the depth model converges when the predicted depth value is as close

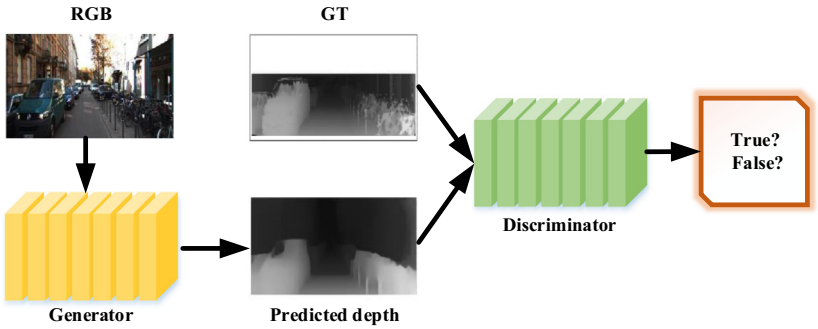



**Fig. 3.** The general GAN-based framework for supervised monocular depth estimation.

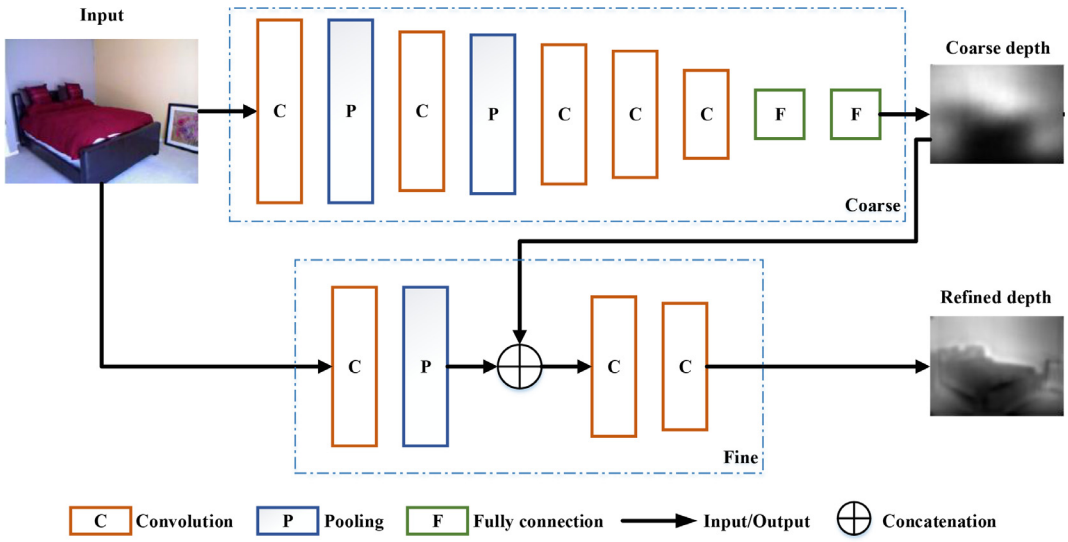



**Fig. 4.** The architecture of multi-scale network for monocular depth estimation proposed by Eigen et al. [29]. The top module is the coarse network for coarse estimation and the bottom module is the fine network for refined depth map.

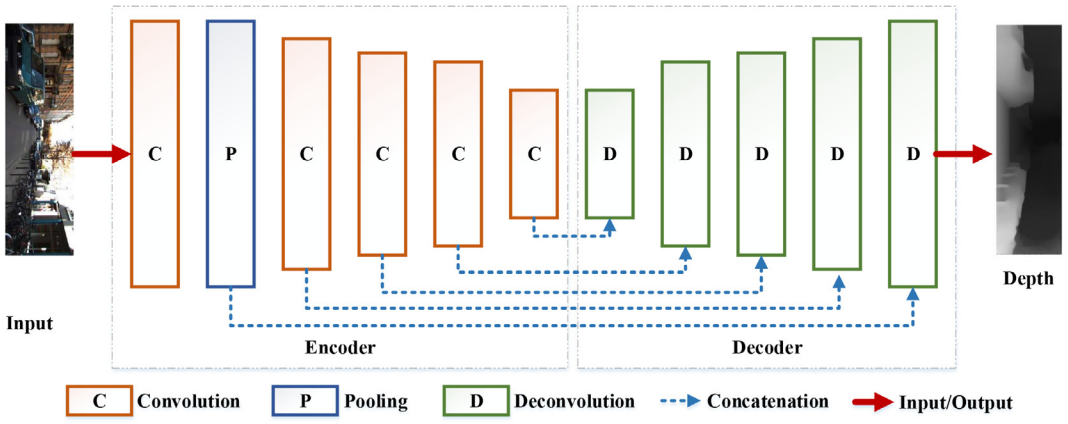



**Fig. 5.** The general pipeline of deep learning for monocular depth estimation. The left module is encoder network learning depth features layer-by-layer, and the decoder network in the right module recovers the depth map.

to GT as possible, and other loss functions are variants of the functions mentioned in Table 1.

*3.1.1.1. CNNs-based methods.* Researchers have designed CNN-based monocular depth estimation networks to learn depth features layer by layer through their convolution kernels and recover depth maps by deconvolution to meet the requirements of scene understanding. This section introduces two aspects based on the absolute depth or relative depth learned from monocular images.

For absolute depth learning, Li et al. [76] proposed a two-streamed framework based on VGG-16 [131] for monocular depth estimation: one stream for depth regression and other for depth gradients, which were combined through a depth-gradient fusion module to obtain a coherent depth map. The entire model was con-

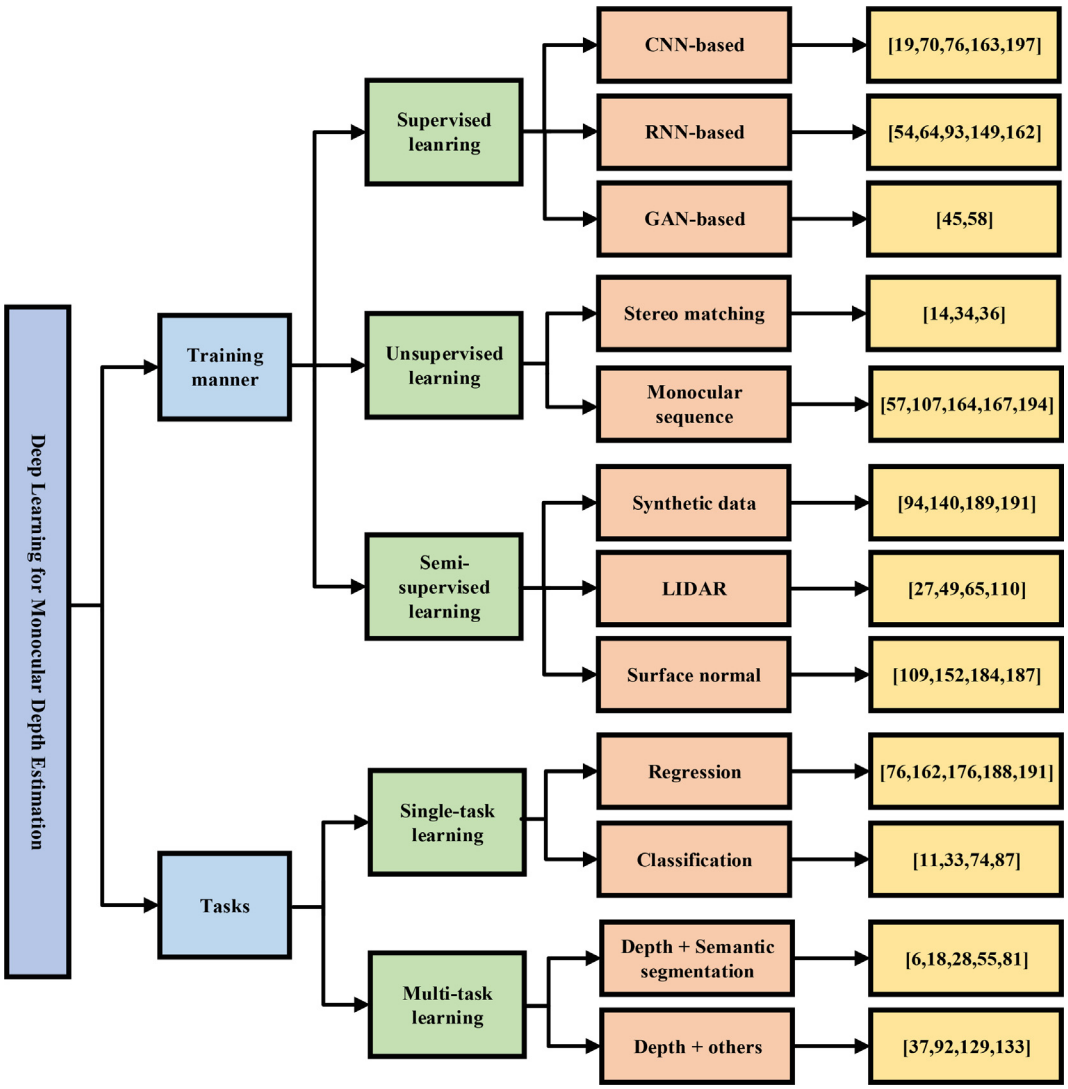



**Fig. 6.** The overall diagram of deep learning methods for monocular depth estimation. According to whether the network is trained with GT, these deep learning methods are divided into supervised, unsupervised, and semi-supervised learning models; according to the types of network prediction task, these methods are classified into single-task and multi-task learning methods.

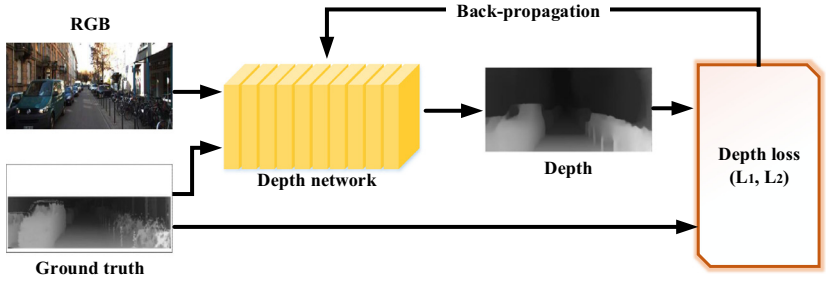



**Fig. 7.** The general model of supervised learning for monocular depth estimation, whose inputs are the RGB and GT depth images and the output is the estimated depth map.

strained by the depth loss and the gradient loss functions, enhancing the generalization abilities of each stream mutually for richer 3D projections. Furthermore, there are many monocular depth estimation methods based on more complex CNNs to learn pixel-level depth, such as VGG-based models [62,188], ResNet-based models [69,71,188], and DenseNet-based models [71].

For relative depth estimation, Zoran et al. [197] proposed a method adopting the relative relationship between point-pairs in the image to infer depth information. They output the relative relationship between the point-pairs and utilized the numerical optimization method to obtain the dense depth maps. Chen et al. [19] proposed a multi-scale network that predicted pixel-level depth by learning relative depth. The network was trained with the relative depth loss function and performed depth recovery on monocular images in an unconstrained environment, whose root mean square error (RMSE) was 1.10 comparable to the absolute

**Table 1**
The loss functions commonly used in supervised learning for monocular depth estimation, where $d$ respects the estimated depth, $d^*$ is the GT depth, $y_i^2 = log(d) - log(d^*)$, $\lambda$ is a balance factor, and $c$ is a threshold.

| Name | Formulation |
|---|---|
| $L_1(d, d^*)$ | $L_1(d, d^*) = \frac{1}{N}\sum_{i=1}^{N}\|\|d_i - d_i^*\|\|_1$ |
| $L_2(d, d^*)$ | $L_2(d, d^*) = \frac{1}{N}\sum_{i=1}^{N}\|\|d_i - d_i^*\|\|_2^2$ |
| $L(logd)$ | $L(d, d^*) = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \frac{\lambda}{N}\left(\sum_{i=1}^{N} y_i\right)^2$ |
| Berhu | $L_{Berhu}(d, d^*) = \begin{cases} \|d - d^*\| & if \quad \|d - d^*\| \leqslant c, \\ \frac{\|d-d^*\|^2 + c^2}{2c} & if \quad \|d - d^*\| > c. \end{cases}$ |

depth estimation model [83]. Lee et al. [70] designed a CNN to estimate the relative depth at different scales, which was optimally reorganized to reconstruct the final depth map. Their RMSE was better than most absolute depth methods mentioned above.

The absolute depth learning has higher accuracy, and the relative depth learning models are more robust which aren't affected by the data homography.

*3.1.1.1.1. Combined with CRF.* Conditional Random Field (CRF) is a conditional probability distribution model under the condition of a given input sequences [66]. CRF can establish a structured connection between input and output, where the key is to construct a reasonable and correct feature for monocular image depth estimation. In order to regress continuous depth, depth estimation networks with fixed and shared weights are constructed to learn different patches firstly. Then, these estimations are propagated to the CRF module to obtain the final depth, as shown in Fig. 8.

Based on CRF, Xu et al. [163] proposed an attention model to automatically learn robust multi-scale features through an integrated attention mechanism [85,146,155], where the cascade-CRFs module reduced the RMSE of 0.088 compared to the baseline based on ResNet-50. Ricci et al. [119] proposed two deep models for monocular depth estimation, one was based on multiple CRF cascading, and the other was based on a unified graph model. Multi-scale features were merged through CRF integration multi-level cascade. Additionally, there are lots of CNNs combined with continuous CRF [75,83], hierarchical CRF [151], FC–CRF [11,100], to predict monocular depth in a supervised manner.

CNN has made great progress in monocular depth estimation recently. On the one hand, it learns and fits deep features to reconstruct the scene depth maps by designing deeper and more complex networks; on the other hand, it combines with CRF to analyze and optimize the predictions of the deep networks, to obtain refined depth map. How to reconstruct the novel networks to adapt to monocular depth estimation is an important research direction.

*3.1.1.2. RNNs-based methods.* RNN-based supervised learning networks for monocular depth estimation capture the spatial features and temporal information from monocular image sequences [54,149]. Different from CNN-based models, the encoder of RNN-based network is designed with all LSTM (or ConvLSTM) layers or consists of convolution and LSTM (ConvLSTM) layers to extract and reserve spatial–temporal features for monocular depth estimation, as shown in Fig. 9.

Kumar et al. [64] proposed the DepthNet with ConvLSTM [162] layers to predict monocular depth maps and implicitly learned the smooth temporal variation. The encoder of DepthNet only consisted of eight ConvLSTM layers likes Fig. 9(a), which made the network fully use the temporal information in sequences, and the convolution operation helped to maintain the spatial geometric relationships between the cells. Furthermore, Mancini et al. [93] adopted LSTM units to exploit the input stream sequentiality and predict scene depth, where the LSTM layers followed the convolution layers in the encoder network, illustrated in Fig. 9(b).

*3.1.1.3. GANs-based methods.* GAN-based supervised networks can generate depth maps close to the GT [45,58], as shown in Fig. 3. Specially, Jung et al. [58] introduced GANs to the monocular depth estimation, where the generator consisted of a GlobalNet to extract global features and a RefinementNet to estimate local structures from the input image. The entire model was trained with an adversarial loss built on the estimated depth map and the GT depth map:

$$\min_G \max_D \mathbb{E}_{x \sim P_{GT}}[logD(x)] + \mathbb{E}_{x^* \sim P_G}[log(1 - D(x^*))] \quad (1)$$

where $G$ is the generator function, $D$ is the discriminator function, $x$ is the depth estimated by the generator, $x^*$ is the GT depth map, and $P$ represents the domain of pixel.

*Summary.* Supervised deep learning methods have been widely studied and applied in monocular depth estimation, mainly including CNN-based, RNN-based and GAN-based models, where the CNN mainly learns the spatial features of the scene, the RNN learns the temporal information from the video sequences, and GAN is introduced to generate and discriminate depth maps. Because the supervised learning methods need plenty of GT depth maps as the supervision, the accuracy rate is high when scale of the

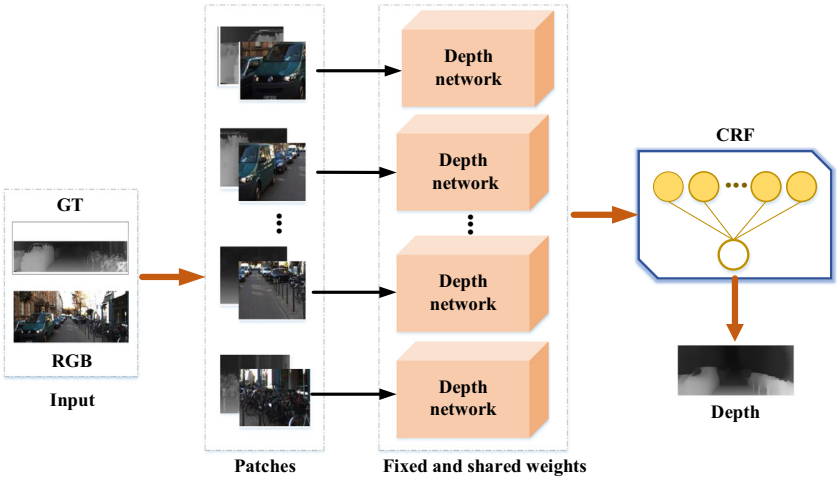



**Fig. 8.** The general model of supervised methods with CRF for monocular depth estimation, where each depth network with fixed and shared weights learns from each pair of patches.

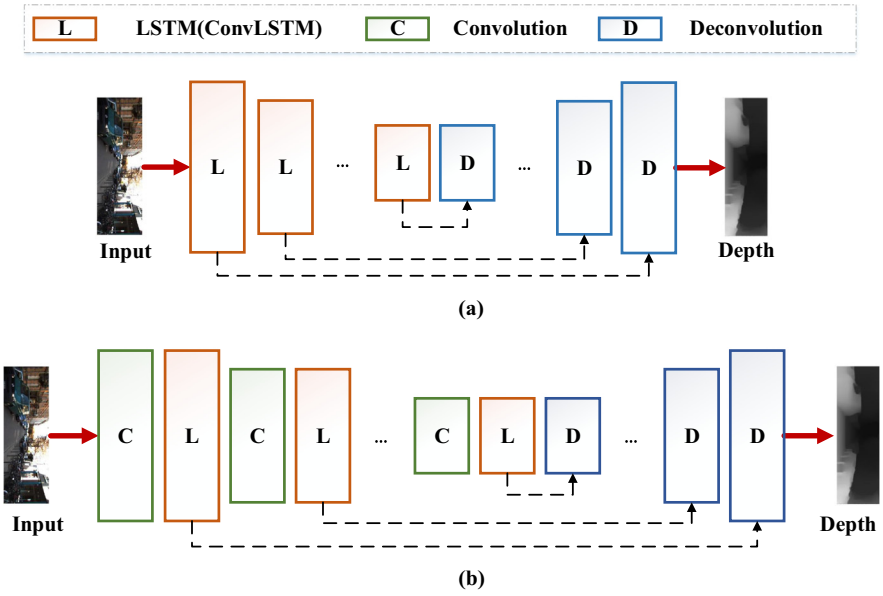



**Fig. 9.** There are two general architectures of RNN-based methods for monocular depth estimation. In (a), the encoder is constructed by all LSTM (or ConvLSTM) layers, yet (b) is composed of convolution and LSTM (or ConvLSTM) layers.

predicted depth map is close to the GT depth map. They can effectively map the 3D structure of the scene. However, GT depth maps are difficult to obtain. Therefore, depth estimation methods based on virtual images have attracted many researchers, and many unsupervised learning methods have emerged, which do not require GT and reduce the requirements for datasets with GT.

*3.1.2. Unsupervised learning methods*

Supervised learning methods need to input a large number of images with GT depth maps during the training stage. However, high-resolution publicly labeled datasets still need numerous equipments and intensive labor work. Therefore, researchers explore unsupervised deep learning methods for monocular depth estimation without GT depth maps. Unsupervised monocular depth estimation are usually trained with stereo pair-wise images or monocular image sequences, and tested on monocular images or sequences, which are trained with scene geometric constraints.

*3.1.2.1. Stereo matching.* Unsupervised learning methods are inspired by traditional stereo matching methods as shown in Fig. 10, which usually utilize left and right images to calculate depth value [136]. The learning model is trained with stereo pair-wise images and tested on single image, as shown in Fig. 11. The depth network estimates the disparity map between the left and right images, where the new image can be constructed with image warping based on the disparity map and the right image. The pixel $p(s)$ can be obtained through

$$p(s) \sim KT(t \to s)D(t)K^{-1}p(t) \tag{2}$$

where $K$ is the camera intrinsics matrix, $T(t \to s)$ is the transformation between left and right images, $D(t)$ is the estimated depth map, and $p(t)$ is the homogeneous coordinate of a pixel in the reconstructed image.

Therefore, the depth network is constrained by the difference, a reconstruction error, between the source and the reconstructed image. Common image reconstruction loss functions are $L_1$ and *SSIM* [156] as follow:

$$L_{rec} = \sum_{p} |I(p) - I^{w}(p)|_1 \tag{3}$$

$$L_{rec} = \alpha \frac{1 - SSIM(I(p) - I^{w}(p))}{2} + (1 - \alpha)|I(p) - I^{w}(p)|_1 \tag{4}$$

where $I(p)$ and $I^{w}(p)$ represents the source image and the warped image reconstructed from the source image, respectively. $\alpha$ is a weight between $L_1$ norm and *SSIM* term.

Unsupervised learning methods based on stereo matching usually adopt CNNs for monocular depth estimation. Garg et al. [34] adopted the general model as shown in Fig. 11 to learn monocular depth maps in an unsupervised manner with the reconstruction loss in $L_1$ norm as Eq. (3) in 2016. On this basis, a number of researchers began to utilize the left and right views to train networks with stereo matching based on 2D CNNs and 3D CNNs.

For 2D CNNs, Godard et al. [36] proposed the left–right consistency constraints to train the unsupervised network, where they reconstructed the left and right view simultaneously. Their model was constrained by the reconstruction loss, the disparity smoothness loss, and the left–right disparity consistency. Experiments proved that the addition of the new loss functions enhanced the accuracy of the predicted depth map from each view. Moreover, Xie et al. [161] added a selection layer in image reconstruction, Wong et al. [158] designed a global-to-local network for feature extraction, Goldman et al. [38] constructed a Siamese network to learn stereo images, Andraghetti et al. [3] enhanced the depth estimation with traditional visual odometry. Watson et al. [157] strengthened stereo matching with depth hints. Ur et al. [115] applied unsupervised pre-trained filter method.

For 3D CNNs, some researchers adopted context information to constrain unsupervised networks in 3D convolution blocks for monocular depth estimation [14,59,60], as shown in Fig. 12. During training, two 2D CNNs with shared weights learn feature maps from left and right images, respectively. And then, these two groups of feature maps are concatenated to the 3D convolution network in a cost volume module [15,143] to estimate the final depth map combined with context information [42,175]. Specially, Chang et al. [14] proposed the PSMNet, trained in a top-down/bottom-up manner to perform unsupervised monocular depth estimation, where a spatial pyramid pooling module was used as a matching cost volume by aggregating semi-global environment information and a 3D convolution module adjusted the matching

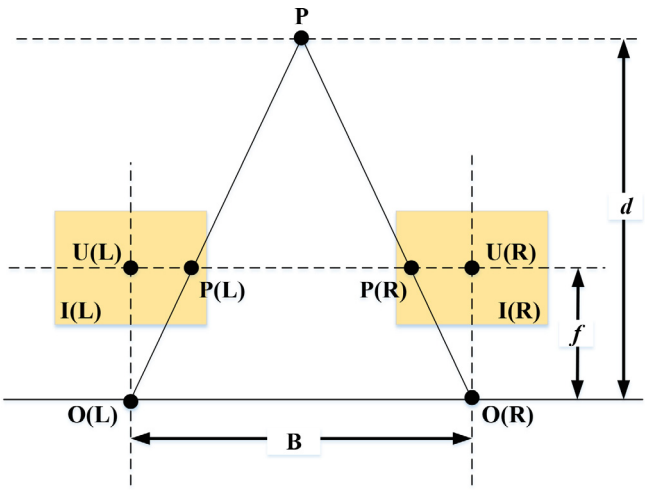



**Fig. 10.** The principle of stereo matching methods for depth estimation, where $I(L)$ and $I(R)$ are stereo pair-wise images taken by the left and the right cameras, respectively.

cost volume by combining multiple stacked hourglass-based 3D CNNs with intermediate supervision.

Unsupervised learning models based on stereo matching is mainly constrained by the projection and mapping relationship between the left and right pair-wise images, which still require the datasets containing stereo images. Therefore, how to utilize only a single camera in the training stage for unsupervised monocular depth estimation has attracted the attention of researchers.

*3.1.2.2. Monocular sequences.* Unsupervised learning models trained with monocular sequences consider the scene structure and camera motion at the same time, where camera pose estimation is similar to the images transformation estimation and has a positive impact on monocular depth estimation [168,190,195]. Recently, researchers have introduced the visual odometry [105,125] into the depth estimation based on monocular sequences, where the scene depth can be learned by predicting the camera motion.

The general model of unsupervised learning based on monocular sequences for depth estimation is shown in Fig. 13, which consists of two sub-networks, depth network for depth estimation and pose network for visual odometry, respectively. During the training stage, these two networks are trained jointly, and the entire model is constrained by image reconstruction loss similar to stereo matching methods. The difference is that the image warping is built on adjacent frames of the monocular sequence. For loss functions, the smoothness loss and the photometric consistency loss in stereo matching methods are adopted in the unsupervised methods based on monocular sequences apart from the reconstruction loss.

Zhou et al. [194] designed two networks to estimate depth maps and camera motion in the monocular video independently, which could be trained jointly or separately with reconstruction loss and photometric consistency loss functions [144,154] and tested on one image or monocular sequence. Their work provided many useful references for subsequent works, such as, models trained with 3D geometric constraints [91,167,193], estimation with uncertainty or confidence maps [16,107], networks designed with self-attention [57], and others [2,164,176].

*Summary.* Unsupervised learning methods for monocular depth estimation directly learn depth information from geometric constraints. It mainly includes two types: one is based on the stereo matching, where the geometric constraints are built on the left and right images; the other is based on monocular sequences, where the geometric constraints are built on adjacent frames. Compared with the supervised learning methods, unsupervised learning methods don't need GT depth maps, which reduces the cost of building depth labels yet suffer from lower accuracy.

### 3.1.3. Semi-supervised learning methods

In order to effectively utilize a large amount of relatively cheap unlabeled data to improve learning performance, researchers have proposed the semi-supervised learning methods, which introduces other information, such as synthetic data, surface normals, and LIDAR, as the semi-supervised learning manners to reduce the model's dependence on GT depth maps, which enhance the scale consistency and improve estimated accuracy of depth maps.

*3.1.3.1. Combined with synthetic data.* The synthetic data generated by the graphics engine provides a possible solution for collecting a large amount of depth data. Thus, researchers introduce synthetic datasets with depth labels to monocular depth estimation. How to overcome the domain gap between synthetic and real data is a challenge during training [10,118].

With the development of image style transfer and its connection with domain adaptation, researchers adopted the style transfer and adversarial training to estimate depth maps in real scenes [5,103], which relied on the models trained with a large amounts of synthetic data, as shown in Fig. 14. The depth estimation network is trained with synthetic images and corresponding GT depth maps. During the test stage, the trained network is applied directly to predict the depth maps from real RGB images with transfer learning to minimize the gap between the real and synthetic domain.

DispNet [94] was the first network that introduced image style transfer for depth estimation. It utilized a large comprehensive synthetic dataset to train, and fine-tuned the model on the less

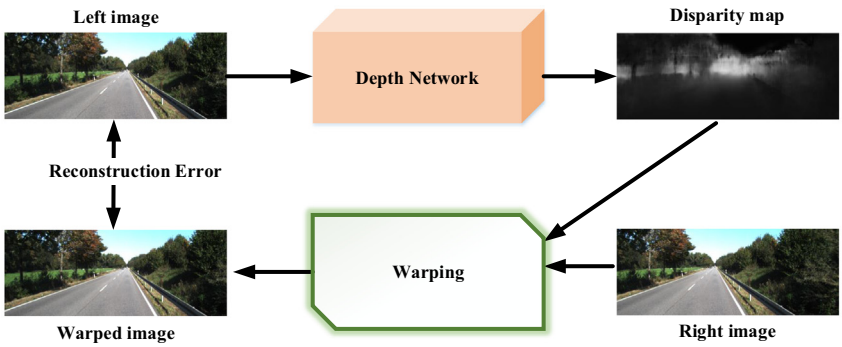



**Fig. 11.** The general model of unsupervised methods with stereo matching for monocular depth estimation.

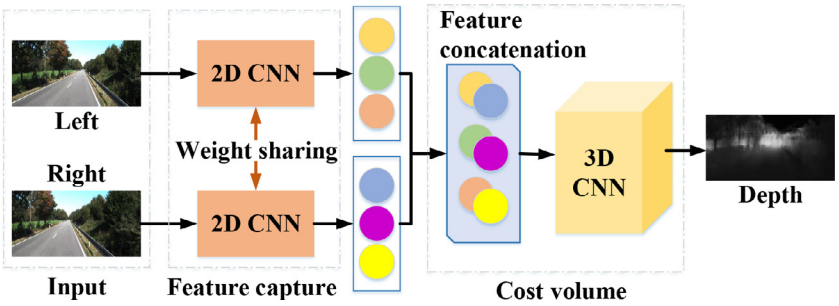



**Fig. 12.** The general model based on unsupervised 2D with 3D CNNs for monocular depth estimation, where the weights of these two 2D CNNs are shared and the cost volume is constrained with context information to mapping the depth map.

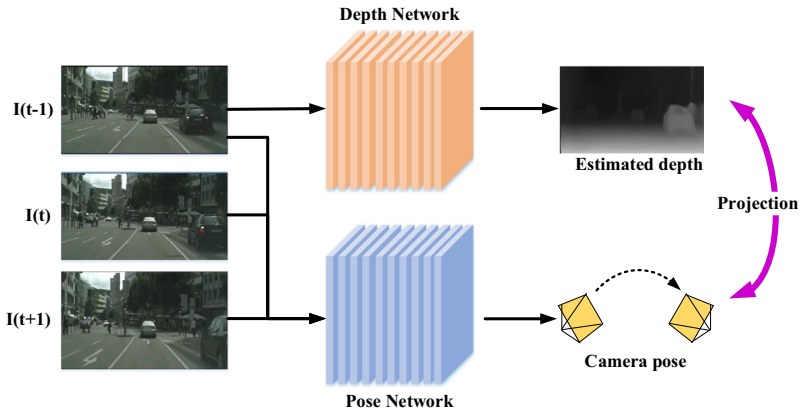



**Fig. 13.** The general model of unsupervised learning based on monocular sequences for depth estimation, where the entire model estimates depth and camera pose simultaneously, and they project and interact with each other.

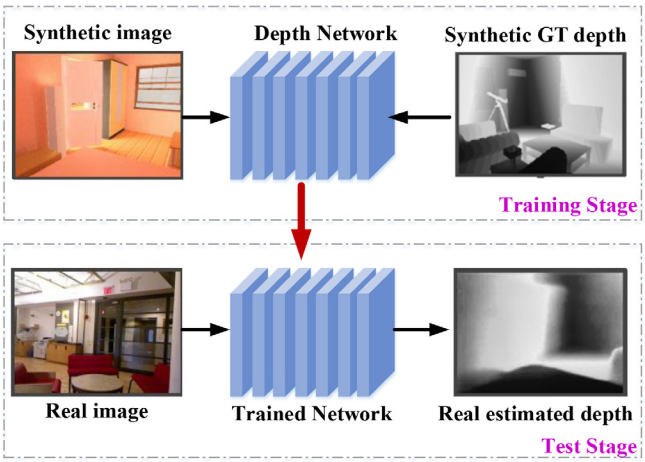



**Fig. 14.** The general model of domain adaptive methods for monocular depth estimation combined with synthetic data, where the network in test stage is trained on synthetic data with GT in training stage.

available GT data. Based on the DispNet, Zheng et al. [192] proposed a two-module domain adaptive network, $T^2$Net, where one module was trained with synthetic and real images and reconstructed each other with the reconstruction loss and generative adversarial loss [21,26,39], and these outputs were input into the other module to predict the real depth maps. Besides, there are more models with self-attention [191], cycle consistency [189], cross-domain [44,140,141], and others for domain adaptation to predict monocular depth maps.

Domain adaptation methods can successfully solve the domain difference of the deep end-to-end disparity estimation network. However, when the illumination or the saturation of the style transfer changes suddenly, the accuracy of the estimated depth map will decrease accordingly.

*3.1.3.2. Combined with LIDAR.* Researchers also adopt auxiliary depth sensors to capture GT information, such as LIDAR, for monocular depth estimation [27,31,49,65,110]. Auxiliary depth sensors cause some noises and the measured depth values are usually sparser than GT depth maps. The general model for monocular depth estimation with LIDAR is shown in Fig. 15. The depth network learns not only structure features but also depth and noise from sparse data captured by LIDAR, where the entire mode needs to add the depth consistency constraint built on the sparse data and estimated depth map as follow:

$$L_{depth}(p) = \sum_{p} ||D(p) - Z(p)||_1 \tag{5}$$

where $p$ is the depth pixel, $D(p)$ is the estimated depth map, and $Z(p)$ is the sparse data from LIDAR.

Kuznietsov et al. [65] proposed a semi-supervised learning network for monocular depth estimation with sparse data, which input left and right images to the model and built a stereo alignment as a geometric constraint. Thus, the depth consistency losses include two parts: one is the error between the left estimated depth map and sparse data, and the other is the error between the right estimated depth map and sparse data. Experiments proved that the added sparse data did improve the performance than supervised and unsupervised methods [28,34,36,83].

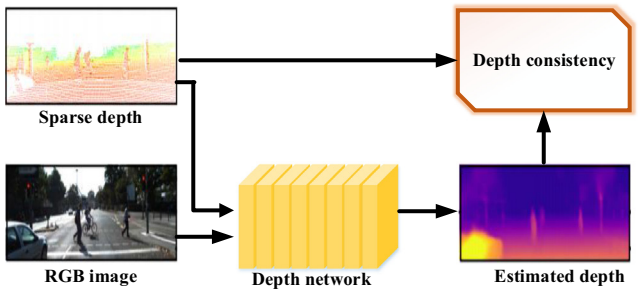



**Fig. 15.** The general model for monocular depth estimation with LIDAR, where the sparse depth is captured by LIDAR.

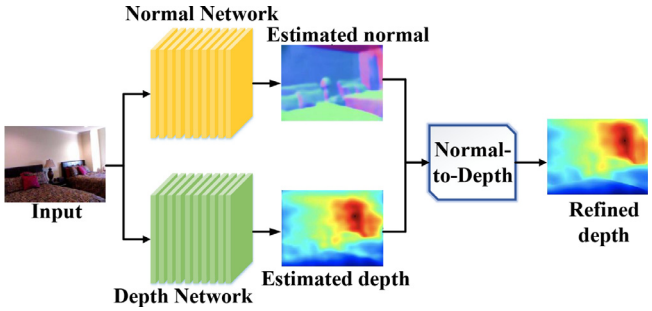



**Fig. 16.** The general model for monocular depth estimation combined with surface normal estimation, where the normal-to-depth module is depended on the geometric relationship between the depth and normal.

*3.1.3.3. Combined with surface normal.* There are still some features with similar information to depth extracted from the input RGB image, which contribute to predict the depth maps more accurately and conveniently, e.g. surface normal.

There is a strong correlation between the surface normal and the depth: the surface normal is determined by the local tangent plane of the 3D point, which can be estimated from the depth; the depth is constrained by the local tangent plane determined by the surface normal. The general model for monocular depth estimation combined with surface normal estimation is shown in Fig. 16. Qi et al. [109] proposed the GeoNet, which consists of a depth-to-normal network exploiting the least square solution of the surface normal from depth and a normal-to-depth network refining the initial depth map in a kernel regression module. They took the advantage of the theory that surface normals change less in local plane to refine monocular depth estimation, where the specific derivation process could be found in Reference [109]. Furthermore, there are some models with depth-normal consistency [110,167], surface regularized constraints [152,187], and depth completion [184], for monocular depth estimation combined with surface normal estimation.

*Summary.* Semi-supervised learning methods for monocular depth estimation relies on auxiliary information, such as virtual data, sparse depth, and surface normals, apart from learning the depth features from the RGB image, which makes the depth map more accurate than that estimated in unsupervised learning methods. Although auxiliary information is easier to obtain than GT depth maps, it still increases the amount of input data and the dependence of depth estimation on it.

### 3.1.4. Summary

This section mainly reviews and summarizes the deep learning methods for monocular depth estimation from the networks training manners, including: supervised, unsupervised, and semi-supervised learning methods. Supervised learning methods for monocular depth estimation have the highest accuracy, yet strong dependence on GT depth maps; unsupervised learning methods build geometric constraints on the input images to predict depth maps without supervision, but its accuracy is slightly inferior to supervised learning and semi-supervised learning methods, where scale ambiguity, occlusion, and other problems need to be overcome; semi-supervised learning methods depend on auxiliary information, which are easier to obtain than GT depth maps. The summaries for different learning manners are concluded in Table 2.

## 3.2. Tasks

From the perspective of task types, deep learning methods for monocular depth estimation can be divided into two categories. On the one hand, we can train a single network only for depth estimation, that is single-task learning; on the other hand, we can combine depth estimation with other related tasks to learn together for the features projection and improve the depth estimation performance, that is multi-task learning. This section will review the two aspects of single-task learning and multi-task learning methods.

### 3.2.1. Single-task learning methods

The core of the single-task learning methods is to construct an association model between the RGB image and the depth map, that is, the model is learned from the RGB image, and recover the depth value. According to whether the depth value returned by the network is continuous or not, single-task learning methods can be divided into regression methods and classification methods.

*3.2.1.1. Regression methods.* Regression methods based on deep learning usually learn scene structure features from inputs and regress continuous depth values to fit the input. Most of the existing monocular depth estimation methods are regression methods, which can directly obtain a depth map containing continuous pixel-level depth values. The general model of regression methods is similar to Fig. 5, where the estimated depth values are continuous.

**Table 2**
A summary of the deep learning methods for monocular estimation in supervised, unsupervised, and semi-supervised learning manners.

| Methods | Models | Descriptions | Remarks | Papers |
|---|---|---|---|---|
| Supervised | Fig. 7 | GT depth maps are used as the supervision signal of the deep learning network. | High precision, simple framework, yet heavy dependence on GT. | [29,33,69,163] |
| Unsupervised | Fig. 11 | Using epipolar geometric constraints instead of GT as the supervision. | GT is not required, but there are problems such as scale blur, dynamic blur, and occlusion. | [34,36,38] |
| Semi-supervised | Figs. 14–16 | Relying on virtual data, sparse depth, surface normal and other auxiliary information. | Heavy dependence on the auxiliary information. | [65,109,192] |

According to the deep learning model used, it can be divided into CNN-based [27,69,76], RNN-based [64,162,176], and GAN-based regression methods [45,191]. Zhang et al. [188] proposed an end-to-end progressive hard mining network (PHN) to regress depth maps, in which an intra-scale module restored the depth information, an inter-scale module fused the depth cues, and a hard-mining refinement module constrained the recursive refining and reduced error propagation to fully learn boundaries of different scales and estimate depth maps in regression.

Ideally, the estimated depth values should be continuous. However, regression methods for monocular depth estimation are usually faced with more complex network structures and constraint functions. Therefore, some researchers began to discretize the depth values and introduced the classification methods to learn depth maps.

*3.2.1.2. Classification methods.* Depth estimation and semantic segmentation are similar, and both are pixel-level predictions. Taking into account the characteristics of the scene from far to near, classification is also used to estimate monocular depth maps, as shown in Fig. 17. Firstly, the continuous depth values are discretized. Then, the depth estimation network learns the corresponding classification labels for discretized depth values and regresses segmented depth maps. Finally, these segmented depth maps are combined into the final depth map.

There are several deep learning models in classification for monocular depth estimation, such as full convolutional models [11], residual models [74,116,134], and ordinal classification models [33,87]. Fu et al. [33] put forward a deep ordered classification network to estimate monocular depth maps. It performed linear sampling on the depth value in logarithmic space, and arranged all categories in descending order according to the distance relationship, where the discrete depth values were used for ordered regression network training. Experiments proved that treating depth estimation as a regression problem might lead to larger errors in areas too far or too close to the camera, while treating as a classification problem could effectively avoid a relatively large error for predicting a larger depth value.

*Summary.* Single-task learning methods for monocular depth estimation mainly include regression and classification methods, where the regression methods directly returns continuous depth values, and the classification methods discretize the depth values firstly and then regress those in piecewise. However, the network and constraint functions of the regression are becoming more and more complex, and it is easy to cause local minima; and the classification method has a strong dependence on the discretization form and weight setting, otherwise the loss will be increased.

*3.2.2. Multi-task learning methods*

In order to make full use of the complementarity of the depth and other features, researchers have proposed to design a unified framework for joint multi-task training, and the features extracted from different tasks are projected to each other to enhance the final depth map. This section introduces the depth estimation methods combined with semantic segmentation in monocular images and the methods combined with visual odometry, optical flow estimation, and others in monocular videos.

*3.2.2.1. Combined with semantic segmentation.* Scene perception includes many aspects, where depth information describes the geometric relationship in space, and the semantic information represents the entity meaning of different parts in the scene [90,106,171]. These tasks share similar context information [24,79]. Many works have been proposed to combine semantic segmentation with depth estimation, processing data under the same neural network [56,98,113,186].

The model for monocular depth estimation and semantic segmentation consists of one encoder network and two decoder networks for depth regression and semantic labels prediction, where these two decoder networks share weights, as shown in Fig. 18. During training, we can train only one or two-both tasks at the same time. The shared encoder learns feature maps from the input, yet two decoders with shared weights to recover depth maps and semantic segmentations, respectively. Furthermore, the whole model is constrained by the attention guidance from context information, and the predicted results will be back-propagation to update network parameters and optimize the results.

Eigen et al. [28] were the first to unify the three tasks of depth, surface normal, and semantic annotation. Based on that, more and more methods have been proposed for monocular depth estimation with semantic segmentation. Atapour-Abarghouei et al. [6] considered depth estimation as a supervised image-to-image translation problem with a generative network and applied adversarial learning to force the model to select a mode to overcome the multi-modal problem resulting in blurry outputs. For semantic segmentation, they applied a fully supervised generative network

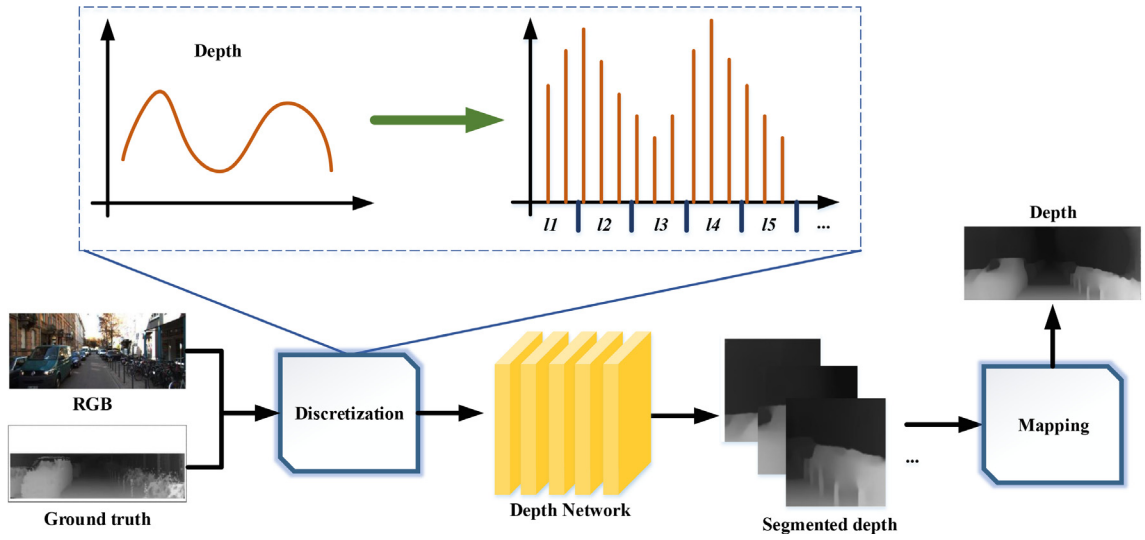



**Fig. 17.** The general model of classification methods for monocular depth estimation, where the discretization module discretizes continuous depth values, and the mapping module combines the segmented depth maps into the final depth map.

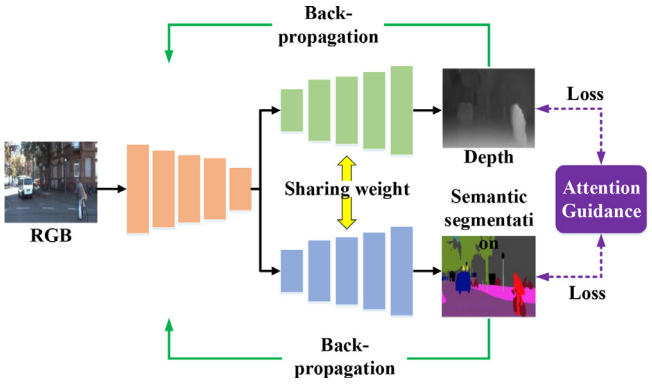



**Fig. 18.** The general model for monocular depth estimation combined with semantic segmentation, where the shared encoder captures the scene structure features and two separate decoders perform semantic segmentation and depth regression respectively.

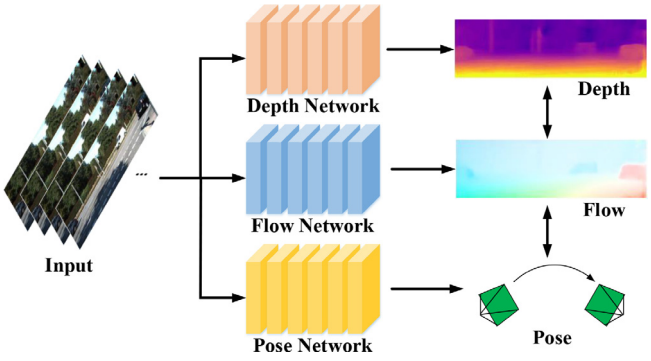



**Fig. 19.** The general model for monocular depth estimation combined with visual odometry and flow estimation, which includes three sub-networks: the depth network, the flow network, and the pose network for depth, scene flow, and camera pose estimation, respectively.

trained with cross-entropy loss functions. What's more, models with self-attention [55], instance segmentation [17,150], multi-scale learning [100], guidance manner [18,43], and others [81,187] are proposed to estimate monocular depth combined semantic segmentation. Experiments proved that the addition of semantic information did increase the accuracy of monocular depth estimation.

Monocular depth estimation combined with semantic segmentation can take advantage of the context information of the scene, overcoming problems such as object boundaries blur and improving the accuracy of the predicted depth maps.

*3.2.2.2. Combined with others.* In addition to combining with semantic segmentation tasks, depth estimation based on monocular video is often combined with other tasks, such as visual odometry [20,30,179] and optical flow estimation [169,173].

Visual odometry is similar to the images transformation estimation and accurate camera pose estimation contributes to image reconstruction and further helps depth estimation [168,190,195]. However, most early methods only consider static scenes, which are no longer applicable in the dynamic scene actually. Because there are usually dynamic objects in real scenes, such as cars and pedestrians. In order to better estimate the depth maps of the dynamic scene, researchers have introduced optical flow estimation into monocular depth estimation. Optical flow estimation can capture motion information in the scene, which contributes to the monocular depth estimation of dynamic scenes [92].

Based on the combination with visual odometry and optical flow estimation, there are a large quantity of works dealing with dynamic objects in the scene and the problems of occlusion and motion blur [37,112,153]. The general model of monocular depth estimation with visual odometry and flow estimation is shown in Fig. 19, which usually consists of multiple sub-networks and each sub-network performs a different task. All tasks are jointly trained and the estimation of each task project and promote each other.

For dynamic objects and occlusion, Godard et al. [37] proposed an automatic occlusion method, Monodepth2, which minimized photometric error to reduce the artifacts at the object boundary, and improved the sharpness of the occlusion boundary. At the same time, they put forward an auto-masking method to filter out some pixels that didn't change in appearance when dynamic objects moved at the same speed as the camera in the scene. Moreover, there are some methods dealing dynamic objects with object masks [147], object motion estimation [12], flow consistency [97,153], displacement field [112], etc.

In addition to combining visual odometry and optical flow estimation, there are some works that combine features estimation for further pixel-level depth maps estimation [129,133,174]. For example, Spencer et al. [133] proposed an unsupervised network framework, DeFeat-Net, that could simultaneously learn monocular depth, dense feature representation, and self-motion. It was robust and could work in many challenging environments, such as changing weather and light conditions, with established pixel-wise loss functions [23,72,132].

*Summary.* Multi-task learning methods for monocular depth estimation usually predict depth maps with other tasks, such as semantic segmentation, visual odometry, and scene optical flow estimation. By capturing features related to depth information in the scene, the accuracy of depth estimation is improved and the scene understanding is enhanced. However, there are still many challenges in multi-task learning that need to be overcome, such as limited datasets with semantic labels or missing labels, motion blur and occlusion caused by dynamic objects in the scene.

### *3.2.3. Summary*

This section mainly reviews and summarizes the deep learning methods for monocular depth estimation based on the task types, including single-task learning and multi-task learning methods. Single-task learning methods usually estimate monocular depth maps in regression or classification manner, distinguished from whether the returned depth values are continuous or discrete. Multi-task learning methods usually combine depth estimation with semantic segmentation, camera pose, and scene flow estimation, which are trained jointly and interact with each other. The summaries for different learning tasks are concluded in Table 3.

## 4. Datasets and metrics

This section introduces the datasets and evaluation metrics of deep learning models for monocular depth estimation.

### *4.1. Datasets*

There are a number of datasets for monocular depth estimation, with different types and depth ranges between indoor and outdoor scenes. This section introduces some common datasets in deep learning methods for monocular deep estimation.

#### *4.1.1. KITTI*

KITTI dataset [35] is an outdoor dataset for monocular deep estimation and object detection and tracking based on deep learning, which is jointly developed by Karlsruhe Institute of Technol-

**Table 3**
A summary of the single-task and multi-task learning methods for monocular estimation, where multi-task learning methods include depth estimation with semantic segmentation.

| Methods | Models | Descriptions | Remarks | Papers |
|---|---|---|---|---|
| Single-task | Fig. 7 | Only perform a single-task of monocular depth estimation. | Predicting monocular depth maps by regression or classification. | [33,87,188] |
| Depth with semantic segmentation | Fig. 18 | Adopting the complementarity between depth information and semantic information for multi-task learning. | The accuracy of depth estimation is improved by applying context information. | [6,18,28,55] |
| Depth with others | Fig. 19 | Using inter-frames geometric constraints and image reconstruction to learn multi-task estimations. | No need for GT, but problems with scale blur, re-projection, dynamic blur, and occlusion. | [37,133,169,194] |

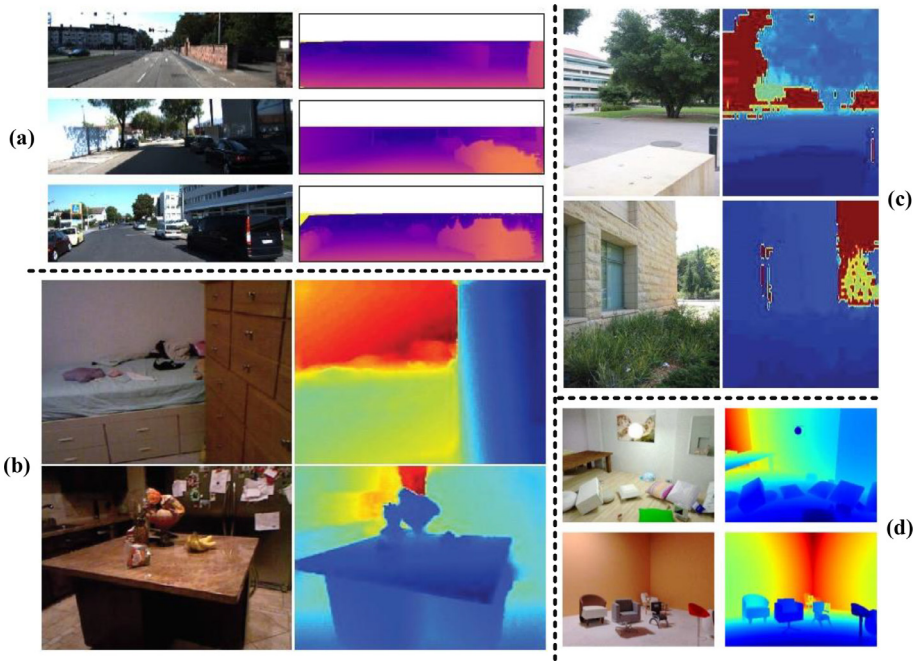



**Fig. 20.** Samples of monocular depth estimation datasets. (a) is KITTI dataset [35], (b) is NYU Depth V2 dataset [130], (c) is Make3D dataset [123,124], and (d) is SceneNet RGB-D dataset [95] (the left images are RGB images and the right are the ground-truth depth maps).

ogy in Germany and Toyota Institute of Technology in the United States, as shown in Fig. 20(a). KITTI dataset is captured through a car equipped with 2 high-resolution color cameras, 2 gray-scale cameras, laser scanner and global positioning system (GPS), whose maximum measuring distance is 120 m. The dataset contains a total of 93,000 RGB-D training samples, including five categories: "Road", "City", "Residential", "Campus", and "Person", from the city of Karlsruhe, the wild area and the highway. The original image size of KITTI is 1,242 × 375, and its ground-truth depth maps are sparse.

*4.1.2. NYU depth V2*

NYU Depth V2 dataset [130] is an indoor dataset for monocular depth estimation based on deep learning, which is provided by Silberman et al. at the New York University. NYU Depth V2 dataset contains 407,024 frames of RGB-D image pairs captured by a Red-Green–Blue (RGB) camera and the Microsoft Kinect depth camera to simultaneously collect the RGB and depth information of 464 different indoor scenes. The original image size of NYU Depth V2 is 640 × 480 and the depth of the dataset ranges from 0.5 m to 10 m. Due to the positional deviation between the RGB and the depth camera, the original depth maps contain missing parts or noises. Therefore, authors select 1,449 images from the dataset and use the coloring algorithm [73] to fill and obtain dense depth maps, which are manually labeled with the semantic information. The 1,449 samples are divided into 795 training samples and 654 test samples. Some samples of NYU Depth V2 dataset are shown in Fig. 20(b).

*4.1.3. Make3D*

Make3D dataset [123,124] is another outdoor dataset for monocular depth estimation based on deep learning, which is constructed by Saxena et al. in Stanford University. Make3D dataset includes daytime city and natural scenery, with depth maps being collected by a laser scanner. The depth ranges from 5 m to 81 m, and the range larger than that is uniformly mapped to 81 m. This dataset contains a total of 534 RGB-D image pairs, 400 of which are used for training and 134 are used for testing. The original resolution of the RGB image is 2,272 × 1,704, and the resolution of the depth map is 55 × 305 pixels. Some samples of Make3D dataset are shown in Fig. 20(c).

*4.1.4. Virtual datasets*

The above datasets, KITTI, NYU Depth V2, and Make3D, are all collected from real scenes. There are some virtual datasets generated by computers, such as SceneNet RGB-D dataset [95], and SYNTHIA dataset [121]. These virtual datasets include various scene types under different weather, environment, and lighting conditions. The appropriate dataset should be selected according

to the specific task in research. Some samples of SceneNet RGB-D dataset are shown in Fig. 20(d).

### 4.2. Metrics

Evaluation metrics proposed by Eigen et al. [29] is adopted to evaluate and compare the performance of depth estimation methods. Evaluation metrics include error and accuracy metrics. The error metrics (smaller is better) include absolute relative error (Abs.rel), square relative error (Sq.rel), root mean square error (RMSE), and the logarithm root mean square error (log RMS); the accuracy rate metrics (the bigger the better) include $\delta < 1.25^t$, where $t$ = 1,2,3. These metrics are formulated as:

$$RMS : \sqrt{\frac{1}{T}\sum_{i\in T}||d_i - d_i^{gt}||^2} \tag{6}$$

$$log\,RMS : \sqrt{\frac{1}{T}\sum_{i\in T}||log(d_i) - log\left(d_i^{gt}\right)||^2} \tag{7}$$

$$abs.relative : \frac{1}{T}\sum_{i\in T}\frac{d_i - d_i^{gt}}{d_i^{gt}} \tag{8}$$

$$sq.relative : \frac{1}{T}\sum_{i\in T}\frac{||d_i - d_i^{gt}||^2}{d_i^{gt}} \tag{9}$$

$$accuracies : \%\ of\ d_i\ s.t.\ max\left(\frac{d_i}{d_i^{gt}}, \frac{d_i^{gt}}{d_i}\right) = \delta < thr \tag{10}$$

where $d_i$ and $d_i^{gt}$ are the predicted and ground-truth depth respectively at the pixel indexed by $i$, and T is the total number of pixels in all the evaluated images.

### 4.3. Analysis and comparisons

In order to evaluate and compare these monocular depth estimation methods based on deep learning, we adopted the publicly available pre-trained networks trained or tested on KITTI [35] dataset. Table 4 illustrates some properties of the deep learning methods, including year, supervision, main contributions, tasks, and training data. The performance comparison of various methods is listed in Table 5, including error metrics and accuracy metrics. We don't describe the properties and performance of all the methods mentioned above, but only summarize some representative models.

## 5. Challenges and trends

Over the past several years, monocular depth estimation based on deep learning has been extensively researched and developed. However, there are still some limitations needed to be overcome.

**Table 4**
Properties of the deep learning methods for monocular depth estimation. "Sup." is "S" representing the supervised, "U" representing the unsupervised, and "Semi" representing the semi-supervised method. "Data" is the training data, where "RGB-D" means RGB and depth maps, "Stereo" means stereo images, "Mono.seq" means monocular sequences, and "Stereo.seq" means stereo sequences.

| Papers | Year | Sup. | Main contributions | Tasks | Data |
|---|---|---|---|---|---|
| Eigen [29] | 2014 | S | Coarse-to-fine, CNN | Depth | RGB-D |
| Eigen [28] | 2015 | S | Multi-scale, CNN. | depth, normal, semantic annotation | RGB-D, semantic labels |
| Zoran [197] | 2015 | S | Relative dense depth, numerical optimization, residual network | Depth | RGB-D |
| Laina [69] | 2016 | S | BerHu loss, residual network | Depth | RGB-D |
| Li [76] | 2017 | S | Two-stream framework, depth-gradient fusion, CNN | Depth | RGB-D |
| Xu [163] | 2018 | S | Cascade-CRFs, attention model | Depth | RGB-D |
| Mancini [93] | 2017 | S | Convolution + LSTM | Depth | Mono.seq + depth |
| Kumar [64] | 2018 | S | ConvLSTM | Depth | Mono.seq + depth |
| Jung [58] | 2017 | S | GAN, global-to-local | Depth | RGB-D |
| Garg [34] | 2016 | U | Image reconstruction, CNN | Depth | Stereo |
| Godard [36] | 2017 | U | Left–right photometric and disparities consistency, disparity smoothness loss | Depth | Stereo.seq |
| Zhou [194] | 2017 | U | Reconstruction and photometric consistency loss | Depth, camera pose | Mono.seq |
| Chang [14] | 2018 | U | Spatial pyramid pooling module, 2D + 3D CNN | Depth | Stereo |
| Zhou [193] | 2018 | U | Bundle adjustment, super-resolution, clip loss | Depth, camera pose | Mono.seq |
| Goldman [38] | 2019 | U | Siamese network, geometric consistency | Depth | Stereo |
| Guizilini [42] | 2020 | U | 3D packing, SfM-based | Depth, camera pose | Mono.seq |
| Poggi [107] | 2020 | U | Depth uncertainty estimation | Depth | Mono.seq |
| Zheng [192] | 2018 | Semi | Domain adaptive, GAN | Depth | Synthetic RGB-D |
| Zhao [189] | 2019 | Semi | Domain adaptive, cycle consistency, GAN | Depth | Synthetic RGB-D |
| Kuznietsov [65] | 2017 | Semi | LIDAR, stereo geometric constraint | Depth | Stereo, sparse GT |
| Qiu [110] | 2019 | Semi | LIDAR, binary mask, attenetion map | Depth, normal | RGB, sparse GT |
| Qi [109] | 2018 | Semi | Normal-to-depth, depth-normal consistency | Depth, normal | RGB |
| Zhang [187] | 2019 | Semi | Cross-task, affinity learning | Depth, normal, semantic segmentation | RGB, semantic labels |
| Dos [27] | 2019 | Semi | Sparse-to-Continuous, Hilbert maps [114], occupancy map | Depth | RGB, sparse GT |
| Zhang [188] | 2018 | S | Progressive hard mining network, learning multi-scale boundaries | Depth | RGB-D |
| Fu [33] | 2018 | S | Ordered regression | Depth | RGB-D |
| Liu [87] | 2020 | S | ConvLSTM, ordinal classification | Depth | Mono.seq |
| Atapour [6] | 2019 | U | Temporally consistency, depth completion, GAN | Depth, flow, semantic segmentation | Mono.seq |
| Wang [150] | 2020 | U | Semantic Divide-and-Conquer Network | Depth, semantic and instance segmentation | Mono.seq |
| Godard [37] | 2019 | U | Per-pixel minimum re-projection and multi-scale estimation for occlusion | Depth, camera pose | Mono.seq |
| Spencer [133] | 2020 | U | Minimun re-projection, auto-masking | Depth, dense feature, camera pose | Mono.seq |

**Table 5**
Evaluation on KITTI dataset and best result is emboldened and bolded. The slower of the error metrics, the better; and the higher of the accuracy metrics, the better. "Sup." is "S" representing a supervised method, "U" representing an unsupervised method, and "Semi" representing a semi-supervised method.

| Methods | Sup. | Abs.rel | Sq.rel | RMSE | log RMS | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
|---|---|---|---|---|---|---|---|---|
| Eigen [29] | S | 0.203 | 1.548 | 6.307 | 0.282 | 0.702 | 0.890 | 0.958 |
| Liu [83] | S | 0.201 | 1.584 | 6.471 | 0.273 | 0.680 | 0.898 | 0.967 |
| Mancini [93] | S | 0.312 | **0.107** | 5.654 | 0.366 | 0.512 | 0.786 | 0.911 |
| Kumar [64] | S | 0.137 | 1.019 | 5.187 | 0.218 | 0.809 | 0.928 | 0.971 |
| Xu [163] | S | 0.122 | 0.897 | 4.677 | - | 0.818 | 0.954 | 0.985 |
| Fu [33] | S | **0.072** | 0.307 | **2.727** | **0.120** | **0.932** | **0.984** | **0.994** |
| Chen [18] | S | 0.118 | 0.905 | 5.096 | 0.211 | 0.839 | 0.945 | 0.977 |
| Garg [34] | U | 0.152 | 1.226 | 5.849 | 0.246 | 0.784 | 0.921 | 0.967 |
| Godard [36] | U | 0.148 | 1.344 | 5.927 | 0.247 | 0.862 | 0.960 | 0.964 |
| Wong [158] | U | 0.133 | 1.126 | 5.515 | 0.231 | 0.826 | 0.934 | 0.969 |
| Goldman [38] | U | 0.113 | 0.898 | 5.048 | 0.208 | 0.853 | 0.948 | 0.976 |
| Andraghetti [3] | U | 0.091 | 0.548 | 3.690 | 0.181 | 0.892 | 0.956 | 0.979 |
| Watson [157] | U | 0.106 | 0.780 | 4.695 | 0.193 | 0.875 | 0.958 | 0.980 |
| Guizilini [42] | U | **0.078** | 0.420 | 3.485 | **0.121** | **0.931** | **0.986** | **0.996** |
| Atapour [6] | U | 0.193 | 1.438 | 5.887 | 0.234 | 0.836 | 0.930 | 0.958 |
| Zhou [194] | U | 0.208 | 1.768 | 6.856 | 0.283 | 0.678 | 0.885 | 0.957 |
| Yin [169] | U | 0.155 | 1.296 | 5.857 | 0.233 | 0.793 | 0.931 | 0.973 |
| Casser [12] | U | 0.109 | 0.825 | 4.750 | 0.1866 | 0.874 | 0.958 | 0.983 |
| Wang [153] | U | 0.112 | **0.418** | **2.320** | 0.153 | 0.882 | 0.974 | 0.992 |
| Godard [37] | U | 0.115 | 0.903 | 4.863 | 0.193 | 0.877 | 0.959 | 0.981 |
| Johnston [57] | U | 0.106 | 0.861 | 4.699 | 0.185 | 0.889 | 0.962 | 0.982 |
| Spencer [133] | U | 0.126 | 0.925 | 5.035 | 0.200 | 0.862 | 0.954 | 0.980 |
| Shu [129] | U | 0.104 | 0.729 | 4.481 | 0.179 | 0.893 | 0.965 | 0.984 |
| Dos [27] | Semi | 0.123 | **0.641** | 4.525 | 0.199 | 0.881 | 0.966 | **0.986** |
| Atapour [5] | Semi | **0.110** | 0.929 | 4.726 | **0.194** | **0.923** | **0.967** | 0.984 |
| Zhao [189] | Semi | 0.143 | 0.756 | **3.846** | 0.217 | 0.836 | 0.946 | 0.976 |
| Zhao [191] | Semi | 0.143 | 0.927 | 4.679 | 0.246 | 0.798 | 0.922 | 0.968 |

1) In order to improve the accuracy, researchers deepen the layers of the deep neural networks, which increases the memory usage and space complexity.
2) In multi-task learning, deep learning methods for monocular depth estimation always apply multiple sub-networks or sub-modules to process different sub-tasks, which increases the amount of calculation and memory consumption.
3) Monocular depth estimation networks usually are encoding-decoding networks. After multiple layers of information processing, the depth features are severely lost, which leads to the low-accuracy estimated depth maps and cannot meet the requirements of practical applications.

In this section, this paper summarizes the key challenges and looks at the directions for future research of monocular depth estimation.

### 5.1. Integration and optimization of the network framework

In many supervised learning models, semantic segmentation will be added with depth estimation, but it is still an independent module that handles independent tasks. In the unsupervised learning methods, there are generally multiple sub-networks which are able to learn depth estimation, visual odometry, and flow estimation, respectively. However, these networks are not well connected, which leads to a large number of parameters increasing the memory requirements and calculations. How to better integrate the network is a research direction and is worth exploring in the future.

We can obtain different features at the same time by using the same deep learning network, such as semantic information, optical flow features, and depth features. In the encoding stage, different types of features are extracted and matched at the same time; in the decoding stage, they are decoded separately to meet the application requirements.

### 5.2. Datasets construction

The quality of datasets largely determines the generalization ability and robustness of the deep learning model. In order to improve the results of depth estimation, more data, with better quality and more scene types, is needed. However, these existing datasets used for depth estimation are relatively limited, and the construction of a new dataset is time-consuming and expensive. At present, some researchers utilize computers to generate a large number of images for depth estimation, but the quality is uneven. How to construct a dataset for monocular depth estimation that meets deep learning is a future research direction.

### 5.3. Dynamic objects and occlusion problems

Realistic scenes are usually complicated, such as containing a large number of moving objects, occlusions, illumination changes, weather changes. However, most of the existing depth estimation models only consider the ideal conditions. Although some researchers have begun to deal with dynamic objects and occlusion scenes and have made some progress recently, how to better estimate the depth of complex scenes to meet practical applications is still a very challenging task, which is an important future research direction.

### 5.4. High-resolution depth map output

Depth estimation is a fundamental step for practical applications such as augmented reality (AR) and virtual reality (VR), and it has a high demand for the accuracy and resolution of the depth map. However, the resolution of the depth predicted by most of the current depth estimation models is usually low, for the purpose of improving calculation efficiency. At present, researchers have used color image super-resolution models [77,80,108] to refine the super-resolution of depth maps [104,120,160]. But how to directly output the high-resolution depth map is still a direction that needs to be studied.

### 5.5. Real-time performance

Image depth estimation is the basic module of SLAM, which is closely integrated with industrial applications, such as autonomous driving. Therefore, practical applications have high requirements for the real-time performance of depth estimation. However, in order to obtain higher accuracy, researchers often construct deeper networks, with more parameters and more constraints, to perform depth estimation, which requires more calculation time and thus cannot meet the real-time requirements of practical applications. Therefore, how to apply a lighter network for real-time estimation while ensuring the accuracy of prediction is a future research direction.

## 6. Conclusion

Monocular depth estimation plays an important role in scene understanding and high-accuracy depth maps are beneficial to the realization of multiple applications. This paper introduces related deep learning models and summarizes deep learning-based monocular depth estimation algorithms, from training manners to task types. Furthermore, this paper also summarizes the properties and performance of these monocular depth estimation methods. Finally, this paper identifies the potential challenges and suggests some future research directions of the monocular depth estimation based on deep learning.

## CRediT authorship contribution statement

**Yue Ming:** Investigation, Formal analysis, Software, Writing - review & editing. **Xuyang Meng:** Investigation, Formal analysis, Software, Writing - review & editing. **Chunxiao Fan:** Resources, Supervision, Writing - review & editing. **Hui Yu:** Methodology, Supervision, Writing - review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


[1] M. Alam, M.D. Samad, L. Vidyaratne, A. Glandon, K.M. Iftekharuddin, Survey on deep neural networks in speech and vision systems, Neurocomputing 417 (2020) 302–321.

[2] Y. Almalioglu, M.R.U. Saputra, P.P. de Gusmao, A. Markham, N. Trigoni, Ganvo: unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 5474–5480.

[3] L. Andraghetti, P. Myriokefalitakis, P.L. Dovesi, B. Luque, M. Poggi, A. Pieropan, S. Mattoccia, Enhancing self-supervised monocular depth estimation with traditional visual odometry, in: 2019 International Conference on 3D Vision (3DV), IEEE, 2019, pp. 424–433.

[4] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, 2017. arXiv preprint arXiv:1701.07875..

[5] A. Atapour-Abarghouei, T.P. Breckon, Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2800–2810.

[6] A. Atapour-Abarghouei, T.P. Breckon, Veritatem dies aperit-temporally consistent depth prediction enabled by a multi-task geometric and semantic scene understanding approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3373–3384.

[7] H. Bay, T. Tuytelaars, L. Van Gool, Surf: speeded up robust features, in: European Conference on Computer Vision, Springer, 2006, pp. 404–417.

[8] A. Bhoi, Monocular depth estimation: a survey, 2019. arXiv preprint arXiv:1901.09402.

[9] A. Bosch, A. Zisserman, X. Munoz, Image classification using random forests and ferns, in: 2007 IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–8..

[10] X. Cao, B. Chen, N. Zeng, A deep domain adaption model with multi-task networks for planetary gearbox fault diagnosis, Neurocomputing 409 (2020) 173–190.

[11] Y. Cao, Z. Wu, C. Shen, Estimating depth from monocular images as classification using deep fully convolutional residual networks, IEEE Trans. Circuits Syst. Video Technol. 28 (2017) 3174–3182.

[12] V. Casser, S. Pirk, R. Mahjourian, A. Angelova, Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8001–8008.

[13] A. Ceni, P. Ashwin, L. Livi, Interpreting recurrent neural networks behaviour via excitable network attractors, Cogn. Comput. 12 (2020) 330–356.

[14] J.R. Chang, Y.S. Chen, Pyramid stereo matching network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5410–5418.

[15] C. Chen, X. Chen, H. Cheng, On the over-smoothing problem of cnn based disparity estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8997–9005.

[16] L. Chen, W. Tang, T.R. Wan, N.W. John, Self-supervised monocular image depth learning and confidence estimation, Neurocomputing 381 (2020) 272–281.

[17] L. Chen, Z. Yang, J. Ma, Z. Luo, Driving scene perception network: real-time joint detection, depth estimation and semantic segmentation, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 1283–1291..

[18] P.Y. Chen, A.H. Liu, Y.C. Liu, Y.C.F. Wang, Towards scene understanding: unsupervised monocular depth estimation with semantic-aware representation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2624–2632.

[19] W. Chen, Z. Fu, D. Yang, J. Deng, Single-image depth perception in the wild, Adv. Neural Inf. Process. Syst. (2016) 730–738.

[20] W. Chen, S. Qian, J. Deng, Learning single-image depth from videos using quality assessment networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5604–5613.

[21] Y. Chen, Y. Zhao, W. Jia, L. Cao, X. Liu, Adversarial-learning-based image-to-image transformation: a survey, Neurocomputing 411 (2020) 468–486.

[22] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014. arXiv preprint arXiv:1406.1078..

[23] B.C. Choy, J. Gwak, S. Savarese, M. Chandraker, Universal correspondence network, Adv. Neural Inf. Process. Syst. (2016) 2414–2422.

[24] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to algorithms, third edition thomas h. cormen, charles e. leiserson, ronald l. rivest, clifford stein, J. Oper. Res. Soc. 42 (2001)..

[25] G.R. Cross, A.K. Jain, Markov random field texture models, IEEE Trans. Pattern Anal. Mach. Intell. (1983) 25–39.

[26] A. CS Kumar, S.M. Bhandarkar, M. Prasad, Monocular depth prediction using generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 300–308..

[27] N. Dos Santos Rosa, V. Guizilini, V. Grassi, Sparse-to-continuous: enhancing monocular depth estimation using occupancy maps, in: 2019 19th International Conference on Advanced Robotics (ICAR), IEEE, 2019, pp. 793–800.

[28] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2650–2658.

[29] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, Adv. Neural Inf. Process. Syst. (2014) 2366–2374.

[30] J.M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, J. Civera, Camconvs: Camera-aware multi-scale convolutions for single-view depth, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2020, pp. 11826–11835.

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

[32] T. Feng, D. Gu, Sganvo: unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks, IEEE Robot. Autom. Lett. 4 (2019) 4431–4437.

[33] H. Fu, M. Gong, C. Wang, K. Batmanghelich, D. Tao, Deep ordinal regression network for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2002–2011.

[34] R. Garg, V.K. Bg, G. Carneiro, I. Reid, Unsupervised cnn for single view depth estimation: geometry to the rescue, in: European Conference on Computer Vision, Springer, 2016, pp. 740–756.

[35] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The kitti vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3354–3361..

[36] C. Godard, O. Mac Aodha, G.J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 270–279.

[37] C. Godard, O. Mac Aodha, M. Firman, G.J. Brostow, Digging into self-supervised monocular depth estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3828–3838.

[38] M. Goldman, T. Hassner, S. Avidan, Learn stereo, infer mono: siamese networks for self-supervised, monocular, depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. (2014) 2672–2680.
[40] A.N. Gorban, E.M. Mirkes, I.Y. Tyukin, How deep should be the depth of convolutional neural networks: a backyard dog case study, Cogn. Comput. (2019) 1–10.
[41] K. Gregor, I. Danihelka, A. Graves, D.J. Rezende, D. Wierstra, Draw: a recurrent neural network for image generation, 2015. arXiv preprint arXiv:1502.04623.
[42] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, A. Gaidon, 3d packing for self-supervised monocular depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2485–2494.
[43] V. Guizilini, R. Hou, J. Li, R. Ambrus, A. Gaidon, Semantically-guided representation learning for self-supervised monocular depth, 2020. arXiv preprint arXiv:2002.12319..
[44] X. Guo, H. Li, S. Yi, J. Ren, X. Wang, Learning monocular depth by distilling cross-domain stereo networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 484–500.
[45] K. Gwn Lore, K. Reddy, M. Giering, E.A. Bernal, Generative adversarial networks for depth map estimation from rgb video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1177–1185.
[46] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, Ghostnet: more features from cheap operations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1580–1589.
[47] S. Hao, Y. Zhou, Y. Guo, A brief survey on semantic segmentation with deep learning, Neurocomputing 406 (2020) 302–321.
[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[49] L. He, C. Chen, T. Zhang, H. Zhu, S. Wan, Wearable depth camera: monocular depth estimation via sparse optimization under weak supervision, IEEE Access 6 (2018) 41337–41345.
[50] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[51] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: efficient convolutional neural networks for mobile vision applications, 2017. arXiv preprint arXiv:1704.04861..
[52] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[53] W. Huang, J. Cheng, Y. Yang, G. Guo, An improved deep convolutional neural network with multi-scale information for bearing fault diagnosis, Neurocomputing 359 (2019) 77–92.
[54] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2012) 221–231.
[55] J. Jiao, Y. Cao, Y. Song, R. Lau, Look deeper into depth: monocular depth estimation with semantic booster and attention-driven loss, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 53–69.
[56] X. Jiao, Y. Chen, R. Dong, An unsupervised image segmentation method combining graph clustering and high-level feature representation, Neurocomputing 409 (2020) 83–92.
[57] A. Johnston, G. Carneiro, Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4756–4765.
[58] H. Jung, Y. Kim, D. Min, C. Oh, K. Sohn, Depth prediction from a single image with conditional adversarial networks, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 1717–1721.
[59] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, A. Bry, End-to-end learning of geometry and context for deep stereo regression, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 66–75.
[60] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, S. Izadi, Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 573–590.
[61] G. Kim, B. Park, A. Kim, 1-day learning, 1-year localization: long-term lidar localization using scan context image, IEEE Robot. Autom. Lett. 4 (2019) 1948–1955.
[62] Y. Kim, H. Jung, D. Min, K. Sohn, Deep monocular depth estimation via integration of global and local predictions, IEEE Trans. Image Process. 27 (2018) 4131–4144.
[63] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.
[64] A.C. Kumar, S.M. Bhandarkar, M. Prasad, Depthnet: a recurrent neural network architecture for monocular depth prediction, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 283–291..
[65] Y. Kuznietsov, J. Stuckler, B. Leibe, Semi-supervised deep learning for monocular depth map prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6647–6655.
[66] J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, (2001) 282–289.
[67] H. Laga, A survey on deep learning architectures for image-based depth reconstruction, 2019. arXiv preprint arXiv:1906.06113..
[68] Z. Lai, E. Lu, W. Xie, Mast: a memory-augmented self-supervised tracker, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6479–6488.
[69] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab, Deeper depth prediction with fully convolutional residual networks, in: 2016 Fourth international conference on 3D vision (3DV), IEEE, 2016, pp. 239–248.
[70] J. Lee, C.S. Kim, Monocular depth estimation using relative depth maps, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 9729–9738.
[71] J.H. Lee, M.K. Han, D.W. Ko, I.H. Suh, From big to small: multi-scale local planar guidance for monocular depth estimation, 2019. arXiv preprint arXiv:1907.10326..
[72] G. Lei, Y. Xia, D.H. Zhai, W. Zhang, D. Chen, D. Wang, Staincnns: an efficient stain feature learning method, Neurocomputing. 406 (2020) 267–273.
[73] A. Levin, D. Lischinski, Y. Weiss, Colorization using optimization, in: ACM SIGGRAPH 2004 Papers, 2004, pp. 689–694..
[74] B. Li, Y. Dai, H. Chen, M. He, Single image depth estimation by dilated deep residual convolutional neural network and soft-weight-sum inference, 2017. arXiv preprint arXiv:1705.00534..
[75] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, M. He, 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, 2015, pp. 1119–1127.
[76] J. Li, R. Klein, A. Yao, A two-streamed network for estimating fine-scaled depth maps from single rgb images, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3372–3380.
[77] T. Li, X. Dong, H. Chen, Single image super-resolution incorporating example-based gradient profile estimation and weighted adaptive p-norm, Neurocomputing 355 (2019) 105–120.
[78] Z. Li, D. He, F. Tian, W. Chen, T. Qin, L. Wang, T.Y. Liu, Towards binary-valued gates for robust lstm training, in: Proceedings of the International Conference on Machine Learning, 2018, pp. 4662–4671.
[79] T.Y. Lin, M. Maire, S. Belongie, J. Hays, C.L. Zitnick, Microsoft coco: common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[80] B. Liu, D. Ait-Boudaoud, Effective image super resolution via hierarchical convolutional neural network, Neurocomputing 374 (2020) 109–116.
[81] B. Liu, S. Gould, D. Koller, Single image depth estimation from predicted semantic labels, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 1253–1260.
[82] C. Liu, J. Gu, K. Kim, S. Narasimhan, J. Kautz, Neural rgb->d sensing: depth and uncertainty from a video camera, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 10986–10995.
[83] F. Liu, C. Shen, G. Lin, Deep convolutional neural fields for depth estimation from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5162–5170.
[84] F. Liu, S. Zhou, Y. Wang, G. Hou, Z. Sun, T. Tan, Binocular light-field: Imaging theory and occlusion-robust depth perception application, IEEE Trans. Image Process. 29 (2019) 1628–1640.
[85] S. Liu, E. Johns, A.J. Davison, End-to-end multi-task learning with attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2019,, pp. 1871–1880.
[86] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, T.D. Pham, Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation, IEEE Trans. Neural Syst. Rehabil. Eng. 28 (2020) 2325–2332.
[87] Y. Liu, Multi-scale spatio-temporal feature extraction and depth estimation from sequences by ordinal classification, Sensors 20 (2020) 1979.
[88] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, IEEE, 1999, pp. 1150–1157..
[89] H. Luo, Y. Gao, Y. Wu, C. Liao, X. Yang, K.T. Cheng, Real-time dense monocular slam with online adapted depth prediction network, IEEE Trans. Multimedia 21 (2018) 470–483.
[90] H. Lyu, H. Fu, X. Hu, L. Liu, Esnet: Edge-based segmentation network for real-time semantic segmentation in traffic scenes, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1855–1859.
[91] R. Mahjourian, M. Wicke, A. Angelova, 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, 2018, pp. 5667–5675.
[92] M. Mancini, G. Costante, P. Valigi, T.A. Ciarfuglia, Fast robust monocular depth estimation for obstacle detection with fully convolutional networks, in: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2016, pp. 4296–4303.
[93] M. Mancini, G. Costante, P. Valigi, T.A. Ciarfuglia, J. Delmerico, D. Scaramuzza, Toward domain independence for learning-based monocular depth estimation, IEEE Robot. Autom. Lett. 2 (2017) 1778–1785.
[94] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, T. Brox, 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, 2016, pp. 4040–4048.
[95] J. McCormac, A. Handa, S. Leutenegger, A.J. Davison, Scenenet rgb-d: can 5m synthetic images beat generic imagenet pre-training on indoor segmentation?, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp 2678–2687.
[96] G. Melis, T. Kočiskỳ, P. Blunsom, Mogrifier lstm, 2019. arXiv preprint arXiv:1909.01792..
[97] X. Meng, C. Fan, Y. Ming, Y. Shen, H. Yu, Un-vdnet: unsupervised network for visual odometry and depth estimation, J. Electron. Imaging 28 (2019) 063015.
[98] Y. Meng, Y. Lu, A. Raj, S. Sunarjo, R. Guo, T. Javidi, G. Bansal, D. Bharadia, Signet: Semantic instance aided unsupervised 3d geometry perception, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9810–9820.
[99] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014. arXiv preprint arXiv:1411.1784..
[100] A. Mousavian, H. Pirsiavash, J. Košecká, Joint semantic segmentation and depth estimation with deep convolutional networks, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 611–619.
[101] F. Mueller, F. Bernard, O. Sotnychenko, M. Verschoor, M.A. Otaduy, D. Casas, C. Theobalt, Real-time pose and shape reconstruction of two interacting hands with a single depth camera, ACM Trans. Graph. (TOG) 38 (2019) 1–13.
[102] R. Mur-Artal, J.M.M. Montiel, J.D. Tardos, Orb-slam: a versatile and accurate monocular slam system, IEEE Trans. Robot. 31 (2015) 1147–1163.
[103] J. Nath Kundu, P. Krishna Uppala, A. Pahuja, R. Venkatesh Babu, Adadepth: unsupervised content congruent adaptation for depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2656–2665.
[104] M. Ni, J. Lei, R. Cong, K. Zheng, B. Peng, X. Fan, Color-guided depth map super resolution using convolutional neural network, IEEE Access 5 (2017) 26666–26672.
[105] D. Nistér, O. Naroditsky, J. Bergen, Visual odometry, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, IEEE, 2004, pp. 964–971..
[106] S.J. Park, K.S. Hong, S. Lee, Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4980–4989.
[107] M. Poggi, F. Aleotti, F. Tosi, S. Mattoccia, On the uncertainty of self-supervised monocular depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3227–3237.
[108] K. Purohit, S. Mandal, A. Rajagopalan, Mixed-dense connection networks for image and video super-resolution, Neurocomputing 398 (2020) 360–376.
[109] X. Qi, R. Liao, Z. Liu, R. Urtasun, J. Jia, Geonet: geometric neural network for joint depth and surface normal estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 283–291.
[110] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, M. Pollefeys, Deeplidar: deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3313–3322.
[111] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015. arXiv preprint arXiv:1511.06434..
[112] M. Ramamonjisoa, Y. Du, V. Lepetit, Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14648–14657.
[113] P.Z. Ramirez, M. Poggi, F. Tosi, S. Mattoccia, L. Di Stefano, Geometry meets semantics for semi-supervised monocular depth estimation, in: Asian Conference on Computer Vision, Springer, 2018, pp. 298–313.
[114] F. Ramos, L. Ott, Hilbert maps: scalable continuous occupancy mapping with stochastic gradient descent, Int. J. Robot. Res. 35 (2016) 1717–1730.
[115] S. ur Rehman, S. Tu, M. Waqas, Y. Huang, O. ur Rehman, B. Ahmad, S. Ahmad, Unsupervised pre-trained filter learning approach for efficient convolution neural network. Neurocomputing 365 (2019) 171–190..
[116] H. Ren, M. El-Khamy, J. Lee, Deep robust single image depth estimation neural network using scene understanding., in: CVPR Workshops, 2019, pp. 37–45..
[117] J. Ren, A. Hussain, J. Han, X. Jia, Cognitive modelling and learning for multimedia mining and understanding, Cogn. Comput. 11 (2019) 761–762.
[118] J. Ren, A. Hussain, J. Zheng, C.L. Liu, B. Luo, Special issue on recent advances in cognitive learning and data analysis, Cogn. Comput. (2020) 1–2.
[119] E. Ricci, W. Ouyang, X. Wang, N. Sebe, et al., Monocular depth estimation using multi-scale continuous crfs as sequential deep networks, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2018) 1426–1440.
[120] G. Riegler, D. Ferstl, M. Rüther, H. Bischof, A deep primal-dual network for guided depth super-resolution, 2016. arXiv preprint arXiv:1607.08569..
[121] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, A.M. Lopez, The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3234–3243.
[122] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.
[123] A. Saxena, S.H. Chung, A.Y. Ng, Learning depth from single monocular images, Adv. Neural Inf. Process. Syst. (2006) 1161–1168.
[124] A. Saxena, J. Schulte, A.Y. Ng, et al., Depth estimation using monocular and stereo cues, in: IJCAI, 2007, pp. 2197–2203..
[125] D. Scaramuzza, F. Fraundorfer, Visual odometry [tutorial], IEEE Robot. Autom. Mag. 18 (2011) 80–92.
[126] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (1997) 2673–2681.
[127] Y. Shen, S. Tan, A. Sordoni, A. Courville, Ordered neurons: Integrating tree structures into recurrent neural networks, 2018. arXiv preprint arXiv:1810.09536..
[128] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, Learning from simulated and unsupervised images through adversarial training, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2107–2116.
[129] C. Shu, K. Yu, Z. Duan, K. Yang, Feature-metric loss for self-supervised learning of depth and egomotion, 2020, 1–16..
[130] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision, Springer, 2012, pp. 746–760.
[131] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014. arXiv preprint arXiv:1409.1556..
[132] J. Spencer, R. Bowden, S. Hadfield, Scale-adaptive neural dense features: learning via hierarchical context aggregation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6200–6209.
[133] J. Spencer, R. Bowden, S. Hadfield, Defeat-net: general monocular depth via simultaneous unsupervised representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14402–14413.
[134] W. Su, H. Zhang, J. Li, W. Yang, Z. Wang, Monocular depth estimation as regression of classification using piled residual networks, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2161–2169.
[135] J. Sun, Z. Wang, H. Yu, S. Zhang, J. Dong, P. Gao, Two-stage deep regression enhanced depth estimation from a single rgb image, IEEE Trans. Emerg. Top. Comput. (2020).
[136] J. Sun, N.N. Zheng, H.Y. Shum, Stereo matching using belief propagation, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2003) 787–800.
[137] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[138] C. Tang, C. Hou, Z. Song, Depth recovery and refinement from a single image using defocus cues, J. Mod. Opt. 62 (2015) 441–448.
[139] G. Tian, L. Liu, J. Ri, Y. Liu, Y. Sun, Objectfusion: an object detection and segmentation framework with rgb-d slam and convolutional neural networks, Neurocomputing 345 (2019) 3–14.
[140] A. Tonioni, M. Poggi, S. Mattoccia, L. Di Stefano, Unsupervised domain adaptation for depth prediction from images, IEEE Trans. Pattern Anal. Mach. (2019), intelligence.
[141] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, L.D. Stefano, Real-time self-adaptive deep stereo, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 195–204.
[142] Y.M. Tsai, Y.L. Chang, L.G. Chen, Block-based vanishing line and vanishing point detection for 3d scene reconstruction, in: 2006 International Symposium on Intelligent Signal Processing and Communications, IEEE, 2006, pp. 586–589..
[143] S. Tulyakov, A. Ivanov, F. Fleuret, Practical deep stereo (pds): toward applications-friendly deep stereo matching, Adv. Neural Inf. Process. Syst. (2018) 5871–5881.
[144] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, T. Brox, Demon: depth and motion network for learning monocular stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 5038–5047.
[145] J. Valentin, A. Kowdle, J.T. Barron, N. Wadhwa, M. Dzitsiuk, M. Schoenberg, V. Verma, A. Csaszar, E. Turner, I. Dryanovski, et al., Depth from motion for smartphone ar, ACM Trans. Graph. (TOG) 37 (2018) 1–19.
[146] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 5998–6008.
[147] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, K. Fragkiadaki, Sfm-net: learning of structure and motion from video, 2017. arXiv preprint arXiv:1704.07804..
[148] K. Wang, X. Peng, J. Yang, S. Lu, Y. Qiao, Suppressing uncertainties for large-scale facial expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6897–6906.
[149] L. Wang, W. Li, W. Li, L. Van Gool, Appearance-and-relation networks for video classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1430–1439.
[150] L. Wang, J. Zhang, O. Wang, Z. Lin, H. Lu, Sdc-depth: semantic divide-and-conquer network for monocular depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 541–550.
[151] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, A. Yuille, Towards unified depth and semantic prediction from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 2800–2809.

[152] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price, A.L. Yuille, Surge: surface regularized geometry estimation from a single image, Adv. Neural Inf. Process. Syst. (2016) 172–180.
[153] R. Wang, S.M. Pizer, J.M. Frahm, 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, 2019, pp. 5555–5564.
[154] S. Wang, R. Clark, H. Wen, N. Trigoni, Deepvo: towards end-to-end visual odometry with deep recurrent convolutional neural networks, in: 2017 International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 2043–2050.
[155] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 7794–7803.
[156] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600–612.
[157] J. Watson, M. Firman, G. Brostow, D. Turmukhambetov, Self-supervised monocular depth hints, 2019, 2162–2171..
[158] A. Wong, S. Soatto, Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5644–5653.
[159] Y. Xia, H. Yu, F.Y. Wang, Accurate and robust eye center localization via fully convolutional networks, IEEE/CAA J. Autom. Sin. 6 (2019) 1127–1138.
[160] Y. Xiao, X. Cao, X. Zhu, R. Yang, Y. Zheng, Joint convolutional neural pyramid for depth map super-resolution, 2018. arXiv preprint arXiv:1801.00968.
[161] J. Xie, R. Girshick, A. Farhadi, Deep3d: fully automatic 2d-to-3d video conversion with deep convolutional neural networks, in: European Conference on Computer Vision, Springer, 2016, pp. 842–857.
[162] S. Xingjian, Z. Chen, H. Wang, D.Y. Yeung, W.K. Wong, W.C. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: Advances in Neural Information Processing Systems, 2015, pp. 802–810..
[163] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, E. Ricci, Structured attention guided convolutional neural fields for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3917–3925.
[164] D. Yang, X. Zhong, D. Gu, X. Peng, H. Hu, Unsupervised framework for depth estimation and camera motion prediction from video, Neurocomputing 385 (2020) 169–185.
[165] X. Yang, Y. Gao, H. Luo, C. Liao, K.T. Cheng, Bayesian denet: monocular depth prediction and frame-wise fusion with synchronized uncertainty, IEEE Trans. Multimedia 21 (2019) 2701–2713.
[166] X. Yang, H. Luo, Y. Wu, Y. Gao, C. Liao, K.T. Cheng, Reactive obstacle avoidance of monocular quadrotors with online adapted depth prediction network, Neurocomputing 325 (2019) 142–158.
[167] Z. Yang, P. Wang, W. Xu, L. Zhao, R. Nevatia, Unsupervised learning of geometry with edge-aware depth-normal consistency, 2017. arXiv preprint arXiv:1711.03665..
[168] X. Ye, X. Ji, B. Sun, S. Chen, Z. Wang, H. Li, Drm-slam: towards dense reconstruction of monocular slam with scene depth fusion, Neurocomputing 396 (2020) 76–91.
[169] Z. Yin, J. Shi, Geonet: unsupervised learning of dense depth, optical flow and camera pose, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1983–1992.
[170] J. Žbontar, Y. LeCun, Stereo matching by training a convolutional neural network to compare image patches, J. Mach. Learn. Res. 17 (2016) 2287–2318.
[171] N. Zeng, H. Li, Z. Wang, W. Liu, S. Liu, F.E. Alsaadi, X. Liu, Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip, Neurocomputing 425 (2021) 173–180.
[172] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, A.M. Dobaie, Facial expression recognition via learning deep sparse autoencoders, Neurocomputing 273 (2018) 643–649.
[173] M. Zhai, X. Xiang, R. Zhang, N. Lv, A. El Saddik, Optical flow estimation using channel attention mechanism and dilated convolutional neural networks, Neurocomputing 368 (2019) 124–132.
[174] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, I. Reid, 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, 2018, pp. 340–349.
[175] F. Zhang, V. Prisacariu, R. Yang, P.H. Torr, Ga-net: guided aggregation net for end-to-end stereo matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 185–194.
[176] H. Zhang, C. Shen, Y. Li, Y. Cao, Y. Liu, Y. Yan, Exploiting temporal consistency for real-time video depth estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1725–1734.
[177] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5907–5915.
[178] J. Zhang, Q. Su, C. Wang, H. Gu, Monocular 3d vehicle detection with multi-instance depth and geometry reasoning for autonomous driving, Neurocomputing 403 (2020) 182–192.
[179] M. Zhang, X. Ye, X. Fan, W. Zhong, Unsupervised depth estimation from monocular videos with hybrid geometric-refined loss and contextual attention, Neurocomputing 379 (2020) 250–261.
[180] P. Zhang, J. Liu, X. Wang, T. Pu, C. Fei, Z. Guo, Stereoscopic video saliency detection based on spatiotemporal correlation and depth confidence optimization, Neurocomputing 377 (2020) 256–268.
[181] R. Zhang, P.S. Tsai, J.E. Cryer, M. Shah, Shape-from-shading: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 21 (1999) 690–706.
[182] T. Zhang, Y. Yang, Y. Zeng, Y. Zhao, Cognitive template-clustering improved linemod for efficient multi-object pose estimation, Cogn. Comput. (2020) 1–10.
[183] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: an extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.
[184] Y. Zhang, T. Funkhouser, Deep depth completion of a single rgb-d image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 175–185.
[185] Z. Zhang, A flexible new technique for camera calibration, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 1330–1334.
[186] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, J. Yang, Joint task-recursive learning for semantic segmentation and depth estimation, in: European Conference on Computer Vision, Springer, 2018, pp. 235–251.
[187] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, J. Yang, Pattern-affinitive propagation across depth, surface normal and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4106–4115.
[188] Z. Zhang, C. Xu, J. Yang, J. Gao, Z. Cui, Progressive hard-mining network for monocular depth estimation, IEEE Trans. Image Process. 27 (2018) 3691–3702.
[189] S. Zhao, H. Fu, M. Gong, D. Tao, Geometry-aware symmetric domain adaptation for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9788–9798.
[190] W. Zhao, S. Zhang, Z. Guan, H. Luo, L. Tang, J. Peng, J. Fan, 6d object pose estimation via viewpoint relation reasoning, Neurocomputing 389 (2020) 9–17.
[191] Y. Zhao, S. Kong, D. Shin, C. Fowlkes, Domain decluttering: simplifying images to mitigate synthetic-real domain shift and improve depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3330–3340.
[192] C. Zheng, T.J. Cham, J. Cai, T2net: synthetic-to-realistic translation for solving single-image depth estimation tasks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 767–783.
[193] L. Zhou, J. Ye, M. Abello, S. Wang, M. Kaess, Unsupervised learning of monocular depth estimation with bundle adjustment, super-resolution and clip loss, 2018. arXiv preprint arXiv:1812.03368..
[194] T. Zhou, M. Brown, N. Snavely, D.G. Lowe, Unsupervised learning of depth and ego-motion from video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1851–1858.
[195] A.Z. Zhu, L. Yuan, K. Chaney, K. Daniilidis, Unsupervised event-based learning of optical flow, depth, and egomotion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 989–997.
[196] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[197] D. Zoran, P. Isola, D. Krishnan, W.T. Freeman, Learning ordinal relationships for mid-level vision, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 388–396.

**Yue Ming** received the B.S. degree in Communication Engineering, and the M.Sc degree in Human–Computer Interaction Engineering, and Ph.D. degree in Signal and Information Processing from Beijing Jiaotong University, China, in 2006, 2008, and 2013. She worked as a visiting scholar in Carnegie Mellon University, U.S., between 2010 and 2011. Since 2013, she has been working as a faculty member at Beijing University of Posts and Telecommunications. Her research interests are in the areas of biometrics, computer vision, computer graphics, information retrieval, pattern recognition, etc.

**Xuyang Meng** is a PhD student at Beijing University of Posts and Telecommunications. She received her BS degree in engineering from Yanshan University in 2016, and her research interests are computer vision and 3-D reconstruction.

**Chunxiao Fan** is currently a professor and the director of Center for information electronic and intelligence system. She served as a member of ISO/IEC JTC1/SC6 WG9, ASN.1 (since 2006) and Chinese Sensor network working group. She also was elevated to evaluation expert of Beijing Scientific and Technical Academy Awards. Her research interests include Heterogeneous media data analysis, Internet of Things, data mining, communication software and so on. In recent years, she is director of several Nation Science Foundation Project. She has published more than 30 papers in international journals and conferences, authored and edited three books and has authorized several patent for invention.

**Hui Yu** is a Professor with the University of Portsmouth, UK. His research interests include methods and practical development in visual computing, machine learning and AI with the applications focusing on human–machine interaction, multimedia, virtual reality and robotics as well as 4D facial expression generation, perception and analysis. He serves as an Associate Editor for IEEE Transactions on Human–Machine Systems and Neurocomputing journal.