sensors

MDPI

Review

Monocular Depth Estimation Using Deep Learning: A Review

Armin Masoumian [1,2,*], Hatem A. Rashwan [1], Julián Cristiano [1], M. Salman Asif [2] and Domenec Puig [1]

[1] Department of Computer Engineering and Mathematics, University of Rovira i Virgili, 43007 Tarragona, Spain; hatem.abdellatif@urv.cat (H.A.R.); julianefren.cristiano@urv.cat (J.C.); domenec.puig@urv.cat (D.P.)

[2] Department of Electrical and Computer Engineering, University of California, Riverside, CA 92521, USA; sasif@ece.ucr.edu

* Correspondence: masoumian.armin@gmail.com

Abstract: In current decades, significant advancements in robotics engineering and autonomous vehicles have improved the requirement for precise depth measurements. Depth estimation (DE) is a traditional task in computer vision that can be appropriately predicted by applying numerous procedures. This task is vital in disparate applications such as augmented reality and target tracking. Conventional monocular DE (MDE) procedures are based on depth cues for depth prediction. Various deep learning techniques have demonstrated their potential applications in managing and supporting the traditional ill-posed problem. The principal purpose of this paper is to represent a state-of-the-art review of the current developments in MDE based on deep learning techniques. For this goal, this paper tries to highlight the critical points of the state-of-the-art works on MDE from disparate aspects. These aspects include input data shapes and training manners such as supervised, semi-supervised, and unsupervised learning approaches in combination with applying different datasets and evaluation indicators. At last, limitations regarding the accuracy of the DL-based MDE models, computational time requirements, real-time inference, transferability, input images shape and domain adaptation, and generalization are discussed to open new directions for future research.

Keywords: monocular depth estimation; single image depth estimation; deep learning; multi-task learning; supervised, semi-supervised, and unsupervised learning

Citation: Masoumian, A.; Rashwan, H.A.; Cristiano, J.; Asif, M.S.; Puig, D. Monocular Depth Estimation Using Deep Learning: A Review. *Sensors* **2021**, *22*, 5353. https://doi.org/10.3390/s22145353

Academic Editor: Anastasios Doulamis

Received: 10 May 2022
Accepted: 15 July 2022
Published: 18 July 2022

1. Introduction

Indisputable breakthroughs in the field of computational photography have helped the emergence of novel functionalities in the imaging process [1,2]. Many works have been carried out so far in the field of computer vision [3–6]. Depth estimation (DE) is a traditional computer vision task that predicts depth from one or more two-dimensional (2D) images. DE estimates each pixel's depth in an image using offline-trained models. In machine perception, recognition of some functional factors such as the shape of a scene from an image and image independence from its appearance seems to be fundamental [7–9]. DE has great potential for use in disparate applications, including grasping in robotics, robot-assisted surgery, computer graphics, and computational photography [10–15]. Figure 1 schematically illustrates the evaluation trend of DE.

The DE task needs an RGB image and a depth image as output. The depth image often consists of data about the distance of the object in the image from the camera viewpoint [16]. The computer-based DE approach has been under evaluation by various investigators worldwide, and the DE problem has been an exciting field of research. Most successful computer-based methods are employed by determining depth by applying stereo vision. With the progress of recent deep learning (DL) models, DE based on DL models has been able to demonstrate its remarkable efficiency in many applications [17–19]. DE can be functionally classified into three divisions, including monocular depth estimation (MDE), binocular depth estimation (BDE), or multi-view depth estimation (MVDE).

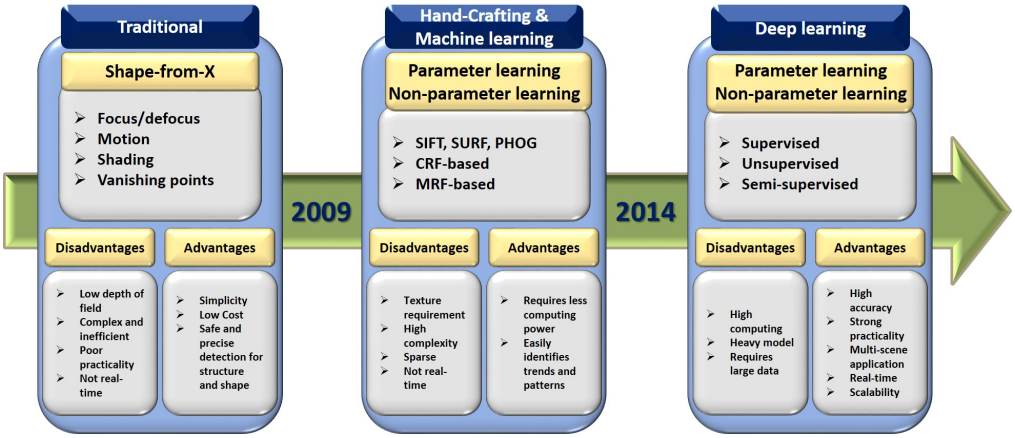

Figure 1. Evaluation trend of DE approaches divided into three sections: traditional methods, hand-crafting and machine learning methods, and deep learning methods.

MDE is an identified significant challenge in computer vision, in which no reliable cues exist to perceive depth from a single image. For instance, stereo correspondences are easily lost from MDE images [20]. Thus, the classical DE methods profoundly depend on multi-view geometry such as stereo images [21,22]. These approaches need alignment procedures, which are of great importance for stereo- or multi-camera depth measurement systems [23,24]. Consequently, using visual cues and disparate camera parameters, BDE and MVDE methods helps to obtain depth information (DI). The majority of BDE or MVDE techniques can accurately estimate DI; however, many practical/operational challenges, such as calculation time and memory requirements for different applications, should be considered [17,25]. The application of monocular images seems to be an excellent idea to capture DI to solve the memory requirement problem. The recent progression in using convolutional neural networks (CNN) and recurrent convolutional neural networks (RNN) yields a considerable improvement in the performance of MDE procedures [26–28].

Scientists worldwide have conducted various medical-based investigations to study the difference in depth perception with MDE or BDE systems. Despite the efforts to use BDE or MVDE systems to estimate depths up to hundreds of meters, the majority of results imply that the most efficient distance for a BDE system is restricted to almost 10 m [29–31]. Small baseline of stereo pairs is the main reason behind the small depth range. Beyond this amount, human vision follows a monocular situation [31]. According to this information, it is obvious that the MDE systems can make better depth predictions than a human. Some problems, including the requirement for a great amount of training data and domain adaptation issues, exist and must be solved appropriately [32].

In addition, research shows that industrial companies are looking at reducing costs and increasing the performance of their AI-based systems. Therefore, this article discusses the main advantages of MDE compared to stereo-based DE due to the low cost of grabbing sensors. In addition, it compares the MDE models from different aspects such as input data shapes and training manner. It discusses the advantages and disadvantages of each model to make it easier for the companies to better understand the differences between these models and select the suitable model for their system.

This paper aims to review the highlighted studies on the recent advancements in the functional application of deep-learning-based MDE. Thus, many DE works from different aspects, including data input types (mono-sequence [16,33,34], stereo sequence [7,26] and sequence-to-sequence [35,36]) and the training manner (i.e., supervised learning (SL) [9,37,38], unsupervised learning (UL) [16,39,40], and semi-supervised learning (SSL) [26,41,42] approaches) combined with the application of different datasets and evaluation indicators have been studied. Eventually, key points and future outlooks such as the accuracy, computational time, resolution quality, real-time inference, transferability, and input data shapes are discussed to open new horizons for future research.

This survey includes over 150 papers, most of them recent, on a wide variety of applications of DL in MDE. To identify relevant contributions, PubMed was queried for papers containing ("Depth Estimation" OR "Relative Distance Prediction") in the title or abstract. ArXiv was searched for papers mentioning one of a set of terms related to computer vision. Additionally, conference proceedings for CVPR and ICCV were searched based on the titles of papers. We checked references in all selected papers and consulted colleagues. The papers without reported results are excluded. When overlapping work had been reported in multiple publications, only the publication(s) deemed most important were included.

Several surveys concerning MDE have been published in recent years, as summarized in Table 1. In this survey, we are concerned with six parameters that are used to assess any MDE method; "TM": training manner, "ACC": accuracy, "CT": computational Time, "RQ": resolution quality, "RTI": real-time inference, "TRAN": transferability, "IDS": input data shapes. In Table 1, we also compare our paper to the recent surveys in terms of the six parameters to show that all of these surveys do not focus on all of these parameters.

Table 1. Comprehensive to the related recent surveys in MDE in terms of six parameters; "TM": training manner, "ACC": accuracy, "CT": computational time, "RQ": resolution quality, "RTI": real-time inference, "TRAN": transferability, "IDS": input data shapes.

Title	Year	TM	ACC	CT	RQ	RTI	TRAN	IDS
Deep-Learning-Based Monocular Depth Estimation Methods [17]	2020	✓	✓	✓				
Monocular Depth Estimation Based on Deep Learning [43]	2020	✓	✓			✓	✓	
Deep Learning for Monocular Depth Estimation [15]	2020	✓			✓			
Towards Real-Time Monocular Depth Estimation for Robotics [44]	2021	✓	✓	✓		✓		
Outdoor Monocular Depth Estimation [45]	2022	✓				✓		
Ours	2022	✓	✓	✓	✓	✓	✓	✓

This survey is organized in the following way: Section 2 describes the background of DE. The DE task's main datasets and evaluation metrics are reviewed in Sections 3 and Section 4, respectively. MDE based on DL models and a comparison of three main data input shapes and training manner approaches are described in Sections 5 and 6. Section 7 presents the discussion, and Section 8 concludes this review.

2. Depth Estimation (DE)

Objects' depth in a scene possesses the remarkable ability of estimation/calculation by applying passive and active approaches. In the active approaches (i.e., applications of LIDAR sensors and RGB-D cameras), the DI is achieved quickly [46,47]. RGB-D camera is a specific type of depth-sensing device that combines an RGB image and its corresponding depth image [48]. RGB-D cameras can be used in various devices such as smartphones and unmanned aerial systems due to their low cost and power consumption [49]. RGB-D cameras have limited depth range and they suffer from specular reflections and absorbing objects. Therefore, many depth completion approaches have been proposed to mitigate the gap between sparse and dense depth maps [44].

In passive techniques, DI is often achieved using two principal methodologies: depth from stereo images and monocular images. The main purpose of both techniques is to assist in building the spatial structure of the environment, which presents a 3D view of the scene. After achieving DI, the situation of the viewer would be recognized relative to the surrounding objects. Stereo vision is a widely-applied depth calculation procedure in the

computer vision area. Stereo vision is known as a computer-based passive approach in which stereo images are applied to extract DI [50–52]. To compute disparity, pixel matching must be implemented among the pixels of both images. It is worth noting that a good correspondence (pixels) matching needs the rectification of both images. Rectification is defined as the transformation process of images to match the epipolar lines of the original images horizontally [53,54]. Figure 2 demonstrates the images before and after the rectification process. The matching process of the pixel in an image with its similar pixel in another image along an epipolar line occurs using a matching cost function. By matching the pixels of both images, the calculation of depth applying the distance between two cameras and the pixel distance between matched pixels will be possible [55,56]. Reflective and highly transparent zones accompanied by smooth areas are the major challenges for stereo matching algorithms. Owing to perspective alteration, an image's edge details can disappear in the second image. If the algorithm does not have sufficient capability to match the edge points on another image, it can create an erroneous depth value and noise in the predicted depth map at those points [57,58].

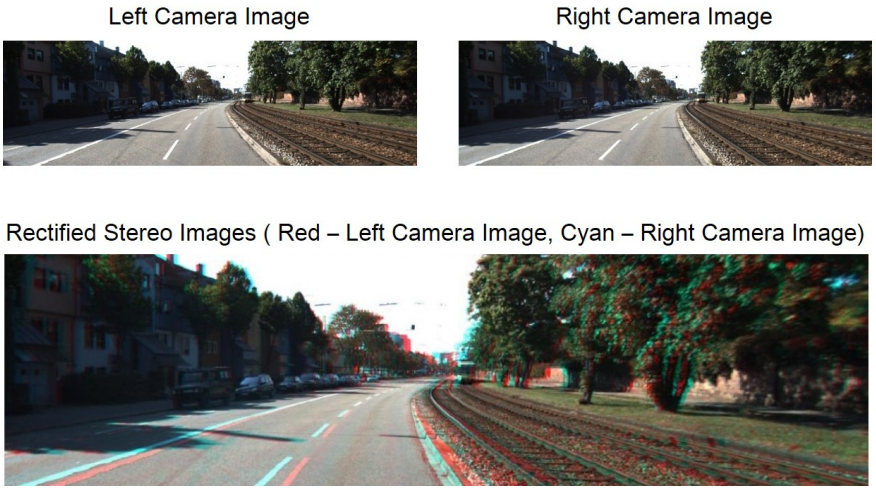

Figure 2. (**Top**) Non-rectified left and right images, and (**down**) red–cyan anaglyph from stereo pair of rectified stereo images.

Sometimes, the application of algorithms for calculating depth may create different challenges. For instance, the matching cost function utilized in the algorithm can generate false-positive signals, which eventuates in the creation of depth maps with low accuracy. Thus, the use of post-processing approaches (i.e., median filter, bilateral filter, and interpolation) is of great importance in stereo vision applications to delete noise and refine depth maps [59–62].

On the contrary, MDE does not require rectified images since MDE models work with a sequence of images extracted from a single camera. This simplicity and easy access are one of the main advantages of MDE compared to stereo models, which require additional complicated pieces of equipment. Because of that, in recent years, demand for MDE increased significantly. Most MDE methods focused on estimating distances between scene objects and the camera from one viewpoint. It is essential for regressing depth in 3D space in MDE methods since there is a lack of reliable stereoscopic visual relationship in which images adopt a 2D form to reflect the 3D space [15]. Therefore, MDE models try to recover the depth maps of images, which reflects the 3D structure of the scene. Most of the MDE models have the main architecture, which contains two main parts: depth and pose networks. The depth network predicts the depth maps. In turn, the pose network works as an ego-motion estimation (i.e., rotation and translation of the camera) between two successive images. The estimated depth (i.e., disparity) maps with the ego-motion parameters used to reconstruct an image should be compared to the target image. Figure 3 represents the schematic illustration of this method.

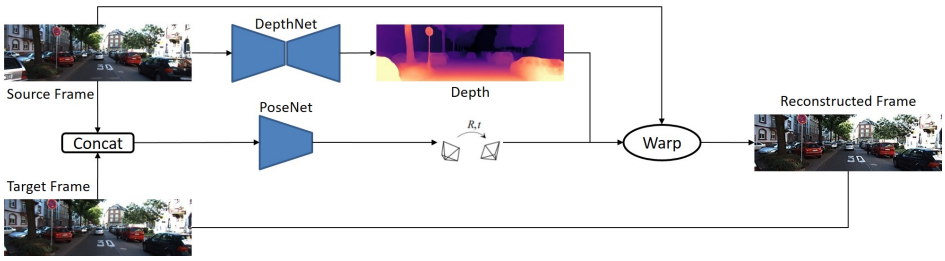

Figure 3. Main network structure for MDE [34]. This network contain two sub-networks: DepthNet for predicting the depth map and PoseNet for estimating the camera pose.

3. Datasets

There are various types of datasets for depth prediction based on different viewpoints. This section highlights the most popular public datasets of DL models for MDE.

3.1. KITTI

The KITTI dataset [63] is considered the most commonly applied dataset in computer vision, such as optical flow, visual odometry (VO), and semantic segmentation [63–66]. This dataset is also the most prevalent criterion in the unsupervised/semi-supervised MDE. In this dataset, 56 scenes are divided into two main compartments: 28 scenes for training and the rest for testing [9]. Due to the incredible capability of the KITTI dataset to create the pose ground truth for 11 odometry sequences, it is extensively applied to assess deep-learning-based VO algorithms [67,68]. This dataset contains 39,810 images for training, 4424 for validation, and 697 for testing. The resolution of the images is 1024 × 320 pixels. The MDE results of the UL, SL, and SSL procedures investigated on the KITTI dataset are presented in Table 2.

Table 2. Comprehensive information about the quantitative results of the SL, SSL, and UL algorithms investigated on the KITTI dataset.

		Lower Better				Higher Better		
Method	**Training Pattern**	**Abs-Rel**	**Sq-Rel**	**RMSE**	**RMSE-Log**	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Bhat [69]	SL	0.058	0.190	2.360	0.088	0.964	0.995	0.999
Wang [70]	SL	0.088	0.245	1.949	0.127	0.915	0.9984	0.996
Patil [71]	SL	0.102	0.655	4.148	0.172	0.884	0.966	0.987
BTS [72]	SL	0.059	0.241	2.756	0.096	0.956	0.993	0.998
DepthNet [35]	SL	0.137	1.019	5.187	0.218	0.809	0.928	0.971
Kuznietsov [73]	SL	0.122	0.763	4.815	0.194	0.845	0.957	0.987
Monodepth [7]	SSL	0.148	1.344	5.927	0.247	0.803	0.922	0.964
SemiSup [26]	SSL	0.113	0.741	4.621	0.189	0.803	0.960	0.986
GMS [74]	SSL	0.143	2.161	6.526	0.222	0.850	0.939	0.972
GAN [75]	SSL	0.119	1.239	5.998	0.212	0.849	0.940	0.976
DepthGAN [76]	SSL	0.152	1.388	6.016	0.247	0.789	0.918	0.965
MonoRes [18]	SSL	0.111	0.867	4.714	0.199	0.864	0.954	0.979
Hints [77]	SSL	0.112	0857	4.807	0.203	0.862	0.952	0.978
SfMLearner [16]	UL	0.208	1.768	6.958	0.283	0.678	0.885	0.957
Vid2Depth [33]	UL	0.163	1.240	6.220	0.250	0.762	0.916	0.968
GeoNet [78]	UL	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Struct2Depth [79]	UL	0.141	1.036	5.291	0.215	0.816	0.945	0.979
CC [80]	UL	0.140	1.070	5.326	0.217	0.826	0.941	0.975
LearnK [81]	UL	0.128	0.959	5.232	0.212	0.845	0.947	0.976
DualNet [82]	UL	0.121	0.837	4.945	0.197	0.853	0.955	0.982
Monodepth2 [83]	UL	0.115	0.882	4.701	0.190	0.879	0.961	0.982
FeatDepth [84]	UL	0.104	0.729	4.481	0.179	0.893	0.965	0.984
GCNDepth [34]	UL	0.104	0.720	4.494	0.181	0.888	0.965	0.984

3.2. NYU Depth-V2

The NYU Depth [85] is a vital dataset, which includes 464 indoor scenes that concentrate on indoor environments. Compared to the KITTI dataset, which collects ground truth with LIDAR, this dataset accepts monocular video sequences of scenes and an RGB-D camera's ground truth of depth. The NYU Depth is the main training dataset in the supervised MDE. The indoor scenes are divided into 249 and 215 sections for training and testing. Due to disparate variable frame rates, there is no one-to-one communication between depth maps and RGB images. Intending to arrange the depth and the RGB images, each depth map is related to the nearest RGB image. In addition, due to the discretion of the projection, all pixels do not possess an associated depth value. Therefore, those pixels that do not have depth value are masked within the experiments [28,85]. The resolution of the RGB images in sequences is 640 × 480 pixels. The MDE results of the investigation on the NYU-V2 dataset are presented in Table 3.

Table 3. Comprehensive information about the quantitative results of the DL algorithms investigated on the NYU-V2 dataset.

		Lower Better				Higher Better		
Method	**Training Pattern**	**Abs-Rel**	**Sq-Rel**	**RMSE**	**RMSE-Log**	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
DeepV2D [86]	SL	0.061	0.094	0.403	0.026	0.956	0.989	0.996
VNL [87]	SL	0.113	0.034	0.364	0.054	0.815	0.990	0.993
Fast-MVSNet [88]	SL	0.551	0.980	3.241	0.243	0.816	0.915	0.939
DORN [28]	SL	0.138	0.051	0.509	0.653	0.825	0.964	0.992
BTS [72]	SL	0.110	0.066	0.392	0.142	0.885	0.978	0.994
GASDA [89]	SSL	1.356	1.156	0.963	1.223	0.765	0.897	0.968
DnD [90]	SSL	0.213	0.320	2.360	0.084	0.761	0.889	0.932
DenseDepth [91]	SSL	0.093	0.589	4.170	0.171	0.886	0.965	0.986
SharpNet [19]	UL	0.139	0.047	0.495	0.157	0.888	0.979	0.995
MonoRes [18]	UL	1.356	1.156	0.694	1.125	0.825	0.965	0.967
DepthComple [92]	UL	0.842	0.760	5.880	0.233	0.863	0.921	0.972
Packnet-SfM [93]	UL	2.343	1.158	0.887	1.234	0.821	0.945	0.968
Monodepth2 [83]	UL	2.344	1.365	0.734	1.134	0.826	0.958	0.979

3.3. Cityscapes

This dataset prominently concentrates on semantic segmentation tasks. In this dataset, 5000 fine-annotation images and 20,000 coarse-annotations images exist [66,94]. Cityscapes dataset includes a series of stereo video sequences, which has only the potential of application for the training process of disparate unsupervised DE procedures [78]. The efficiency of depth networks can be significantly improved by pretraining the networks on the Cityscapes [7,16,95]. The training data of this dataset include 22,973 stereo image pairs with a resolution of 1024 × 2048.

3.4. Make3D

These data include both monocular RGB and depth images but do not possess stereo images that are different from the datasets mentioned above [96,97]. Due to the nonexistence of monocular sequences in the Make3D dataset, SSL and UL procedures do not apply it as the training set, while SL techniques often adopt it for training. The fact of the matter is that the Make3D dataset is extensively used as a testing set of unsupervised algorithms to assess the production capability of networks on disparate datasets [7]. The RGB image resolution is 2272 × 1704, and the depth map resolution is 55 × 305 pixels. The MDE results of the investigation on the Make3D dataset are presented in Table 4.

Table 4. Comprehensive information about the quantitative results of the DL algorithms investigated on the Make3D dataset.

Method	Training Pattern	Abs_Rel	Sq_Rel	RMSE	$\log_{10}$
Karsch [98]	SL	0.428	5.079	8.389	0.149
Liu [99]	SL	0.475	6.562	10.05	0.165
Laina [100]	SL	0.204	1.840	5.683	0.084
SfMLearner [16]	UL	0.383	5.321	10.47	0.478
DDVO [101]	UL	0.387	4.720	8.090	0.204
Monodepth2 [83]	UL	0.322	3.589	7.417	0.201
Jia [102]	UL	0.289	2.423	6.701	0.348
GCNDepth [34]	UL	0.424	3.075	6.757	0.107

3.5. DIODE

DIODE [103] is the Dense Indoor/Outdoor Depth dataset for monocular depth estimation comprising diverse indoor and outdoor scenes acquired with the same hardware setup. This dataset consists of 8574 indoor and 16,884 outdoor samples from 20 scans each for training and 325 indoor and 446 outdoor samples with each set from 10 different scans for validation with the resolution of 768 × 1024. The indoor and outdoor ranges for the dataset are 50 m and 300 m, respectively.

3.6. Middlebury 2014

Middlebury [104] is a dense indoor scene dataset which contains 33 images of 6-megapixel high resolution. Images are captured via two stereo DSLR cameras and two point-and-shoot cameras. Disparity ranges are between 200 and 800 pixels at a resolution of 6 megapixels. The image resolution of this dataset is 2872 × 1984.

3.7. Driving Stereo

The driving stereo [105] is one of the new large-scale stereo driving datasets that contains 182k images. The disparity images are captured via LIDAR, the same as the KITTI dataset. They mainly focus on two new metrics, a distance-aware metric and a semantic-aware metric, for evaluating stereo matching on MDE. The image resolution of this dataset is 1762 × 800. Table 5 represents the summary of datasets features for DE.

Table 5. A summary of depth estimation public datasets.

Dataset	Sensors	Annotation	Type	Scenario	Images	Resolution	Year
KITTI [63]	LIDAR	Sparse	Real	Driving	44 K	1024 × 320	2013
NYU-V2 [106]	Kinect V1	Dense	Real	Indoor	1449	640 × 480	2012
Cityscapes [94]	Stereo Camera	Disparity	Real	Driving	5 K	1024 × 2048	2016
Make3D [96]	Laser Scanner	Dense	Real	Outdoor	534	2272 × 1704	2008
DIODE [103]	Laser Scanner	Dense	Real	In/Outdoor	25.5 K	768 × 1024	2019
Middlebury 2014 [104]	DSLR Camera	Dense	Real	Indoor	33	2872 × 1984	2014
Driving Stereo [105]	LIDAR	Sparse	Real	Driving	182 K	1762 × 800	2019

Although many valuable datasets and benchmarks exist for assessing monocular and stereo DE methods, there are still some limitations in the available datasets. For instance, all these datasets include images captured only during day or night, yet there are no datasets to have both together, and the same applies for indoor or outdoor images. In addition, no dataset concerns different challenges related to the change in weather conditions (e.g., fog, sunny, snow, etc.).

4. Evaluation Metrics

To assess the efficiency of the DE models, an accepted evaluation procedure was recommended by Eigen et al. [9], which possesses five evaluation metrics, including absolute relative difference (Abs-Rel), square relative error (Sq-Rel), root mean square error (RMSE), RMSE-log, and accuracy, with a threshold (δt). They are formulated using the following equations [9]:

$$Abs - Rel = \frac{1}{|D|} \sum_{pred \in D} |gt - pred|/gt \tag{1}$$

$$Sq - Rel = \frac{1}{|D|} \sum_{pred \in D} ||gt - pred||^2/gt \tag{2}$$

$$RMSE = \sqrt{\frac{1}{|D|} \sum_{pred \in D} ||gt - pred||^2} \tag{3}$$

$$RMSE - Log = \sqrt{\frac{1}{|D|} \sum_{pred \in D} ||\log(gt) - \log(pred)||^2} \tag{4}$$

$$\delta t = \frac{1}{|D|} |\{pred \in D \, | max(\frac{gt}{pred}, \frac{pred}{gt}) < 1.25^t\}| \times 100\% \tag{5}$$

In these equations, the pred and gt denote predicted depth and ground truth, respectively. D represents the set of all predicted depths value for a single image, $|\,.\,|$ returns the number of the elements in each input set, and δt represents the threshold.

5. Input Data Shapes For MDE Applying Deep Learning

This section mainly introduces common types of data input for MDE. The input data shapes in MDE networks can be divided into three main categories: mono-sequence, stereo sequence, and sequence-to-sequence input data. Based on the architecture of the networks, the input data shapes will be different.

5.1. Mono-Sequence

Monocular sequence input is mainly used for training the UL models. Figure 4 shows the basic structure of mono-sequence models, which have a single input image and a single output image. UL networks consist of a depth network for predicting depth maps and a pose network for camera pose estimation. The camera pose estimation works similarly to image transformation estimation, which helps to improve the results of MDE. These two sub-networks are connected in parallel, and the whole model is obliged to reconstruct the image. In mono-sequence, mostly the geometric constraints are built on adjacent frames. Lately, researchers have used VO [107] to predict the camera motion for learning the scene depth. Zhou et al. [16] were the pioneers of mono-sequence input type, and they proposed a network to predict camera motion and depth maps with photometric consistency loss and reconstruction loss.

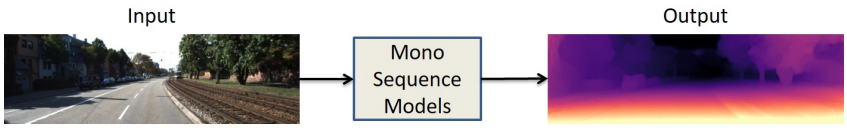

Figure 4. Data input/output structure of mono-sequence models. Single image input and single image output.

Furthermore, Mahjourian et al. [33] introduced a network with 3D geometric constraints and enforced consistency of the estimated 3D point clouds and ego-motion across consecutive frames. Recently, Masoumian et al. [34] designed two jointly connected sub-

networks for depth prediction and ego-motion. They used CNN-GCN encoder–decoder architecture for their networks with three losses: reconstruction loss, photometric loss, and smooth loss. In addition, Shu et al. [84] proposed a similar method with two jointly connected depth and pose predictions that were slightly different. They also added a feature extractor encoder to their model to improve the quality of their predicted depth maps. Their proposed architecture is shown in Figure 5.

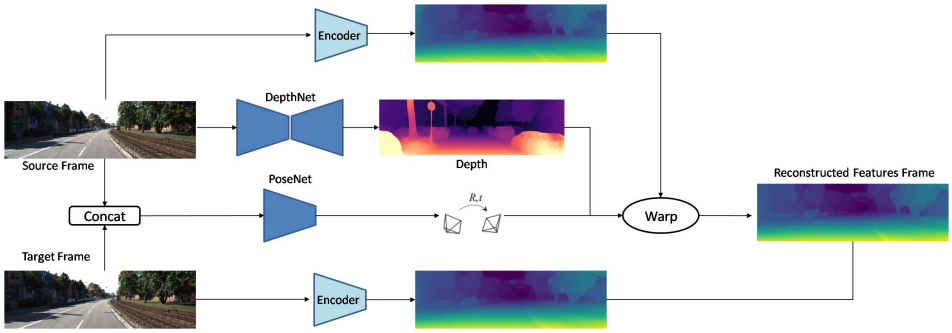

Figure 5. Developed network by Shu et al. [84].

5.2. Stereo Sequence

The projection and mapping relationship between the left and right pairwise images is mainly constrained by stereo matching. In order to build geometric constraints, a stereo images dataset is required. These types of inputs are commonly used in UL and SL networks. Figure 6 represents the basic structure of stereo sequence models which have left and right images as input and a single output. Similar to the monocular sequence input data shape, the stereo sequence works with image reconstruction with slight differences. An image will be reconstructed based on warping between the depth map and the right image. For instance, Kuznietsov et al. [26] proposed an SSL model for MDE with sparse data, and they built a stereo alignment as a geometric constraint.

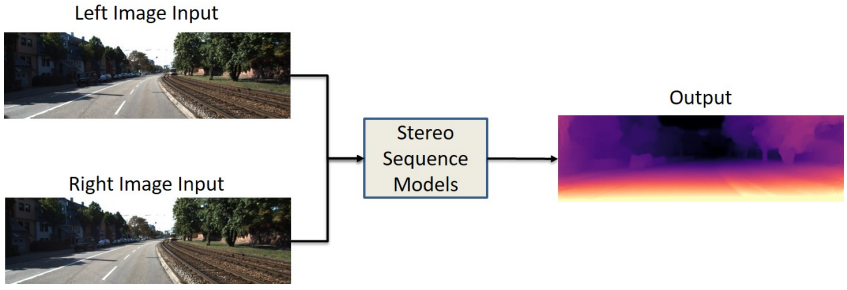

Figure 6. Data input/output structure of stereo sequence models. Stereo pairs of images as an input and single image output.

Furthermore, Godard et al. [7] designed a UL network with left–right consistency constraints. They used CNN-based encoder–decoder architecture for their model with the reconstruction loss, left–right disparity consistency, and disparity smoothness loss. Recently, Goldman et al. [108] proposed a Siamese network architecture with weight sharing, which consists of two twin networks, each learning to predict a disparity map from a single image. Their network is composed of an encoder–decoder pair with skip connections, which is shown in Figure 7.

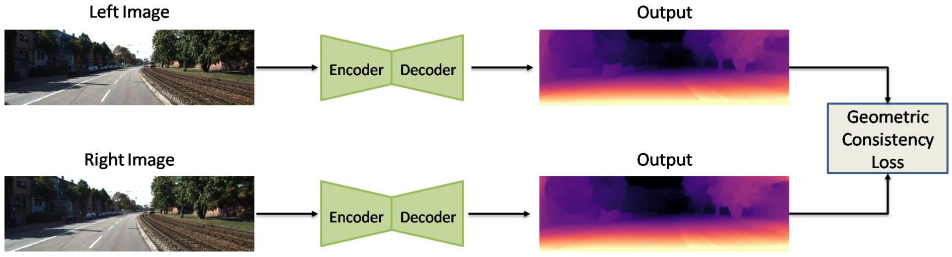

Figure 7. Developed network by Goldman et al. [108].

5.3. Sequence-to-Sequence

Sequence-to-sequence data input is necessary for recurrent neural network (RNN) models [109]. These models have memory capability, which helps the system learn a group of features in sequence images. Figure 8 represents the basic structure of sequence-to-sequence models, which have a sequence of images as input and a sequence of depth maps as an output. Most RNN methods use long short-term memory (LSTM) to learn the long-term dependencies with a three-gate structure [109]. However, RNN and CNN networks will be combined to extract spatial–temporal features. The sequence-to-sequence data primarily will be trained on SL models. Kumar et al. [35] proposed an MDE model with ConvLSTM layers for learning the smooth temporal variation. Their model consists of encoder–decoder architecture, which is shown in Figure 9. Furthermore, Mancini et al. [36] improved LSTM layers to obtain the best outcome of the predicted depth maps by feeding the input images sequentially to the system.

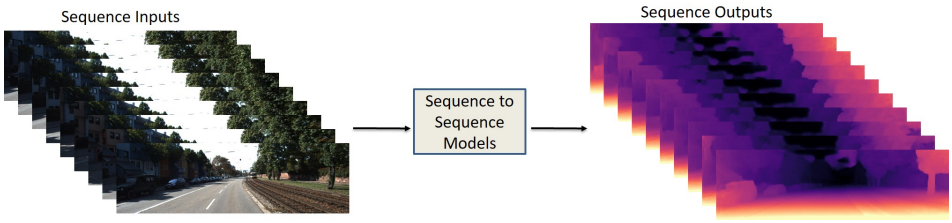

Figure 8. Data input/output structure of sequence-to-sequence models. Sequence of images as an input and sequence of images as an output.

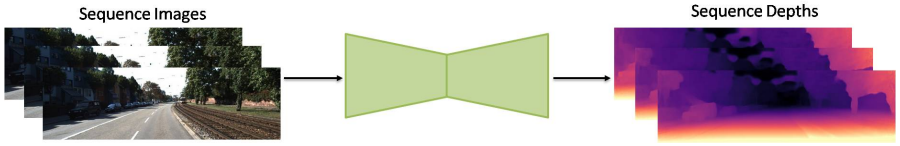

Figure 9. Developed network by Kumar et al. [35].

6. Mde Applying Deep Learning Training Manners

Although DE from multiple images possesses a lengthy background in the computer vision area, the DI extraction process from single images is considered a novel concept in DL. The advancements have initiated comprehensive investigations of the DI concept in DL techniques. The most critical challenge towards the application of DL is the absence of datasets that fit the problem [110–112]. This challenge may also be of great importance for the MDE network. Data applied in training may be collected by LIDAR sensors, RGB-D cameras, or stereo vision cameras. Despite the expensive data collection process, disparate learning strategies have been developed to decrease dependency on the dataset used for training. The learning process in MDE networks can be divided into three parts, including SL, UL, and SSL [7,9,26,37,40,113].

6.1. Supervised Learning Approach

The SL approach for DE needs pixel-wise ground truth DI [114]. The SL procedure applies ground truth depth (GTD) to train a neural network as a regression model [83,115,116]. Eigen et al. [9] were pioneers in investigating DI to train a model using DL. They explained that their developed CNN-based network consists of two deep network stacks. Figure 10 presents a schematic illustration of the network structure proposed in [9]. As shown in Figure 10, the preparation of the input image occurred for both stacks. Additionally, the preparation of the output depth map of the first stack takes place to refine the depth map. The main responsibility of the second stack is to arrange obtained coarse depth predictions with the objects in the scene [9].

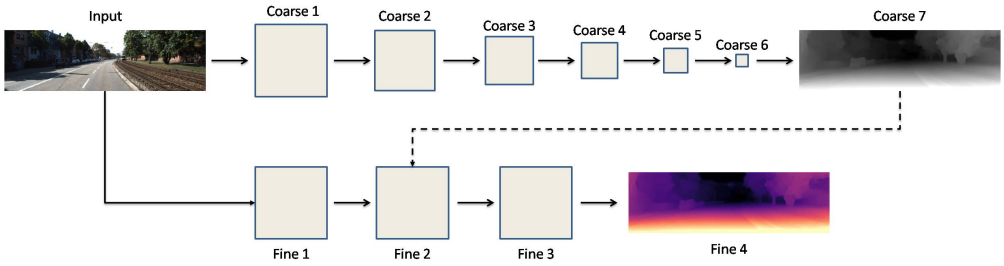

Figure 10. Developed network structure by Eigen et al. [9].

After Eigen's investigation, different procedures were implemented to increase the precision of the estimated depth map (EDP). For example, Li et al. [117] developed a DL network applying conditional random fields (CRFs). They utilized a two-stage network for depth map estimation and refinement. In the first stage, a super-pixel technique on the input image is applied, and image patches are extracted around these super-pixels. In the second stage, CRFs are applied to refine the depth map by changing the super-pixel depth map to the pixel level. In order to extract an appropriate depth map, some approaches use geometric relationships. For example, Qi et al. [37] utilized two networks to estimate the depth map and surface normal from single images. Figure 11 depicts the developed network in [37]. These two networks enable the conversion of depth-to-normal and normal-to-depth and collaboratively increase the accuracy of the depth map and surface normal. Although their neural network can increase the accuracy of depth maps, for training, they require ground truth, including surface normal, which is hard to obtain. Ummenhofer et al. worked on developing a network to estimate depth maps using the structure from motion (SfM) technique. They corroborated that basic encoder–decoder architecture does not have sufficient capacity to process two input images simultaneously. Therefore, they developed a computer-based neural architecture that can extract optical flow, ego-motion, and a depth map from an image pair [38].

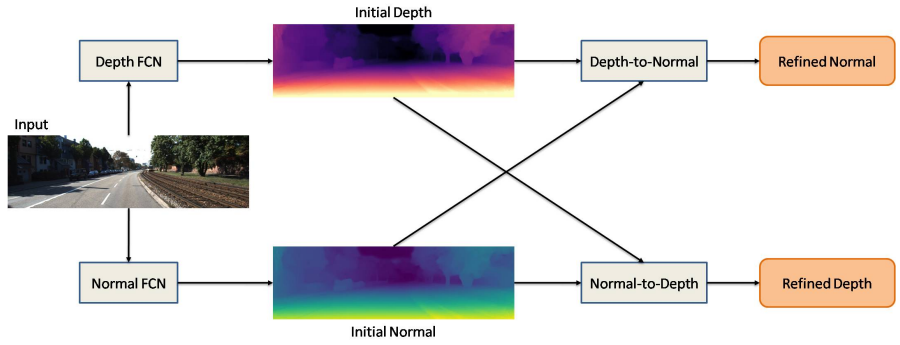

Figure 11. Developed geometric neural network by Qi et al. [37].

The dataset's quality is an introductory section in SL systems, similar to methodology. Dos Santos et al. [118] paid enough attention to this challenge. They developed an approach to creating denser GTD maps from sparse LIDAR measurements via enhancing the valid depth pixels in depth images. They compared the obtained results of their trained model with both sparse GTD maps and denser GTD maps. They understood that the application of denser ground truth results yields increasing performance compared to sparse GTD maps. Ranftl et al. [119] developed an outstanding learning strategy that can involve various datasets to improve the efficiency of the MDE network. To prepare their dataset for three-dimensional movies, they applied stereo matching to conclude the depth of frames of these movies. Disparate unclear problems, including changing resolution and negative/positive disparity values, emerged during the creation of this dataset. According to the assistance of their developed procedures for incorporating multiple datasets, they achieved high precision with their model MDE problem. Recently, Sheng et al. [120] proposed a lightweight SL model with local–global optimization. They used an autoencoder network to predict the depth and used a local–global optimization scheme to realize the global range of scene depth.

6.2. Unsupervised Learning Approach

Increment of layers and trainable parameters in deep neural networks significantly increases the requirement for the train data, resulting in difficulty in achieving GTD maps. For this reason, UL approaches become an appropriate choice because unlabeled data is relatively easier to find [39,121,122]. Garg et al. [40] were the pioneers of developing a promising procedure to learn depth in an unsupervised fashion to remove the requirement of GTD maps. Up until now, developed UL approaches have applied stereo images, and thus, supervision and train loss depend intensely on image reconstruction. In order to train a depth prediction network, consecutive frames from a video may have great potential for application as supervision. Camera transformation estimation (pose estimation) between successive frames is the major challenge of this procedure, which results in extra complexity for the network. As illustrated in Figure 12, Zhou et al. [16] developed computer-based architecture to estimate depth map and camera pose simultaneously. As input, three successive frames are fed to the network. Pose CNN and Depth CNN estimate relative camera poses and a depth map from the first image.

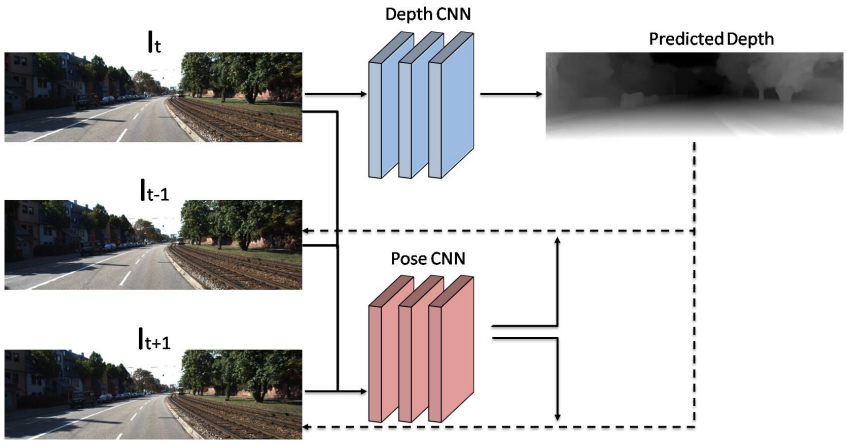

Figure 12. Developed network by Zhou et al. [16].

In order to obtain greater accuracy in DE, some approaches have existed that possess the great potential of application to merge multiple self-supervision procedures into one. For instance, Godard et al. [83] applied MDE and estimated relative camera poses to build other stereoviews and contiguous frames in the video sequence. They added a pose network to their model to predict relative camera pose in adjacent frames. One of the crucial challenges towards using self-supervised approaches via video is occluded pixels. They

applied minimum loss compared to the classical average loss to obtain non-occluded pixels, which is known as a significant improvement [7]. The improvement in the precision of UL approaches has motivated other investigators to modify knowledge distillation methods for the MDE problem. Pilzer et al. developed a system to adapt an unsupervised MDE network to the teacher–student learning framework by applying stereo image pairs to train a teacher network. Despite the promising performance of their student network, it was not as accurate as their teacher network [123]. Masoumian et al. [34] developed a multi-scale MDE based on a graph convolutional network. Their network consists of two parallel autoencoder networks: DepthNet and PoseNet. The DepthNet is an autoencoder composed of two parts: encoder and decoder; the CNN encoder extracts the feature from the input image, and a multi-scale GCN decoder estimates the depth map, as illustrated in Figure 13. PoseNet is used to estimate the ego-motion vector (i.e., 3D pose) between two consecutive frames. The estimated 3D pose and depth map are used to construct a target image.

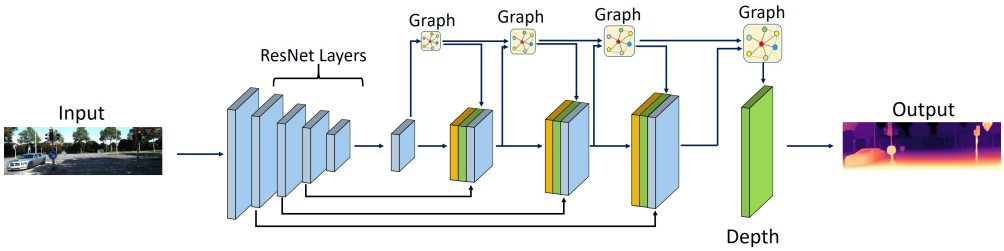

Figure 13. Developed network by Masoumian et al. [34].

6.3. Semi-Supervised Learning Approach

Compared to SL and UL approaches, few investigations have been conducted to study the performance of SSL methods for MDE. Apart from SL and UL approaches, Kuznietsov et al. [26] developed an SSL method by simultaneously applying supervised/unsupervised loss terms during training. Figure 14 demonstrates the components/inputs of the developed semi-supervised loss function in [26].

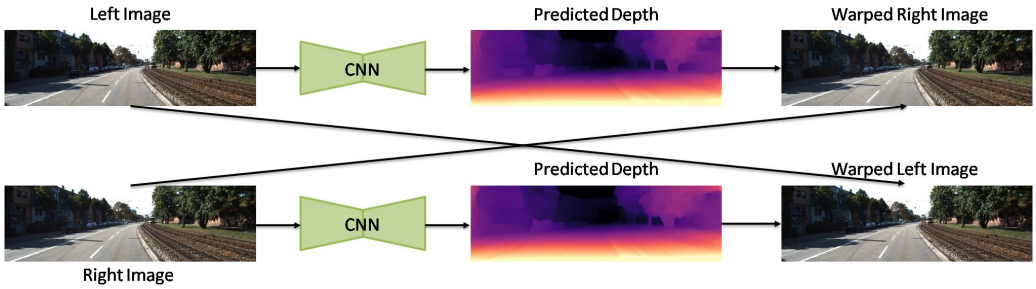

Figure 14. Components/inputs of the developed semi-supervised loss function by Kuznietsov et al. [26].

In their approach, the estimated disparity maps (i.e., inverse depth maps) were used to rebuild left and right images via warping. Computation of unsupervised loss term took place by rebuilding the target images. Simultaneously, the calculation of the supervised loss term occurred by the estimated depth, and GTD maps [26]. Luo et al. [41] classified the MDE problem into two subdivisions and investigated them separately. Based on their procedure, the network requirement for labeled GTD data decreased. Additionally, they corroborated that the application of geometric limitations during inference may significantly increase the efficiency and the performance. Their proposed architecture is shown in Figure 15. Their developed architecture consists of two sub-networks, including view synthesis network (VSN) and stereo matching network (SMN). Their proposed VSN synthesizes the right image of the stereo pair via the left image. In SMN, simultaneous application of left and

synthesized right images occurs in an encoder–decoder architecture pipeline to achieve a disparity map. In SMN, GTD maps are used to calculate the loss for estimated depth maps.

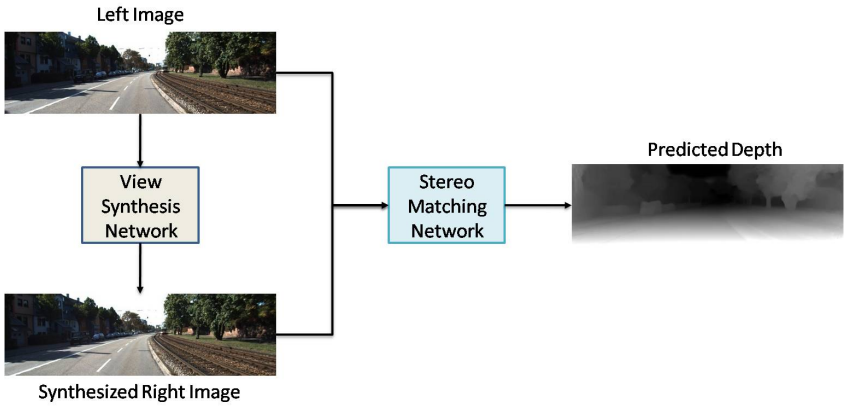

Figure 15. Components/inputs of developed semi-supervised loss function by Luo et al. [41].

Cho et al. [124] developed a novel teacher–student learning strategy to train an MDE network in an SSL approach. Their proposed procedure is demonstrated in Figure 16. They first introduced a stereo matching network with GT labeled data and permitted the teacher network to estimate depth from stereo pairs of an extensive unlabeled dataset. Then, they applied the aforementioned estimated depth maps/unlabeled dataset to train an optimized student network for MDE [124]. They also investigated the trade-off between the precision and the density of pseudo labeled depth maps. The density increases as the pixels in the depth map increase. They concluded the increment of the pseudo labeled depth maps' precision by enhancing the density. Additionally, they reported that their MDE network achieved the greatest accuracy when the density of pseudo labeled depth maps was almost 80% [124].

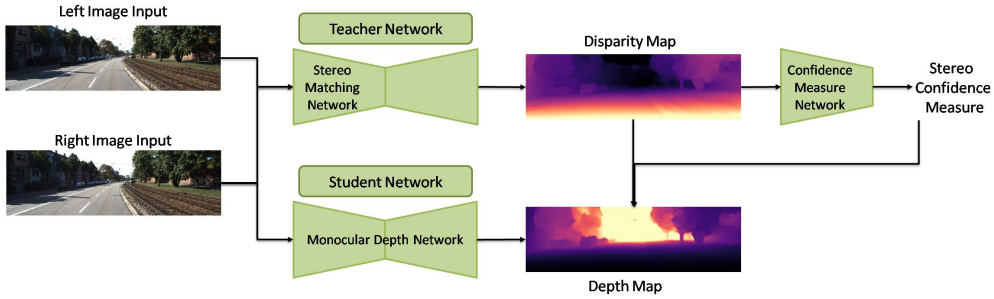

Figure 16. Developed network by Cho et al. [124].

7. Discussion

Due to the ability of humans to use theoretical-based information about the world, estimating depth maps from a single image may be easy for them [43]. Relying on the aforementioned fact, former investigations obtain MDE via mixing some old data, such as the communication between some geometric structures [17,28,125,126]. Due to the acceptable efficacy of image processing, CNN has illustrated a powerful capability to precisely predict dense depth maps from single images [9,127]. In recent years, numerous researchers have studied different types of cues of depth networks required for MDE according to four corroborated procedures, including MonoDepth, SfMLearner, Semidepth, and GCNDepth [7,16,26,34]. Deep neural networks are identified as a black box. In this black box, the supervised signals are applied to accelerate the learning process of some structural information for depth inference. The lack of sufficient datasets with ground truth due to their high economic cost can be considered one of the most critical DL problems.

Table 6 aims to represent comprehensive information about the existing procedures based on their training data, supervised signals, and contributions.

Table 6. Comprehensive information about the applied procedures in the deep learning of MDE.

Ref	Training Set	SL	SSL	UL	Major Contribution
Mousavian et al. [128]	RGB + Depth	✓			Multi-task (semantic + depth)
Jung et al. [129]	RGB + Depth	✓			Adversarial learning, global-to-local
Mayer et al. [64]	RGB + Depth	✓			Multi-task (optical flow + depth)
Laina et al. [100]	RGB + Depth	✓			Residual learning, BerHu loss
Kendall et al. [130]	Stereo sequences + Depth	✓			End-to-end learning
Fu et al. [28]	RGB + Depth	✓			Ordinal regression
Facil et al. [131]	RGB + Depth	✓			Multi-scale convolution
Wofk et al. [132]	RGB + Depth	✓			Lightweight network
Garg et al. [40]	Stereo sequences		✓		Image reconstruction, CNN
Chen et al. [133]	RGB + Relative depth annotations		✓		The wild scene dataset
Godard et al. [7]	Stereo sequences		✓		Left–right consistency
Kuznietsov et al. [26]	Stereo sequences + LIDAR		✓		Direct image alignment
Ramirez et al. [74]	Stereo sequences + Semantic label		✓		Semantic prediction
Pilzer et al. [76]	Stereo sequences		✓		Cycled generative network
Aleotti et al. [75]	Stereo sequences		✓		Generative adversarial network
He et al. [134]	Stereo sequences + LIDAR		✓		Sparse optimization
Fei et al. [135]	Stereo sequences + IMU + Semantic label		✓		Physical information
Li et al. [136]	Stereo sequences		✓		Absolute scale recovery
Zhao et al. [89]	Stereo sequences + Synthesized Depth		✓		Domain adaptation
Wu et al. [137]	Mono-sequences+LIDAR		✓		Attention mechanism
Zhou et al. [16]	Mono-sequences			✓	Ego-motion framework
Wang et al. [138]	Stereo sequences			✓	Multi-task (optical flow + depth)
Zhan et al. [39]	Stereo sequences			✓	Deep feature reconstruction
Chen et al. [139]	Mono-sequences			✓	Connecting flow, depth, and camera
Gordon et al. [81]	Mono-sequences			✓	Camera intrinsic prediction
Li et al. [140]	Mono-sequences			✓	Sequential adversarial learning
Almalioglu et al. [141]	Mono-sequences			✓	Generative adversarial network
Godard et al. [83]	Mono-sequences			✓	Left–right consistency
Shu et al. [84]	Mono-sequences			✓	Feature metric
Masoumian et al. [34]	Mono-sequences			✓	Graph multi-layer

7.1. Accuracy

To achieve high accuracy, several factors are involved. The first factor is using the supervised or unsupervised model. Our evaluation proves that supervised methods achieved higher accuracy than unsupervised and semi-supervised methods due to labeling the original ground truth. However, collecting a large dataset of monocular videos with accurate depth maps is a challenging task. Therefore, we can consider that unsupervised methods perform better than supervised methods if we neglect the slight difference in precision against the time for labeling data. Another factor is the frameworks of the developed networks. For instance, developing a DL model, such as graph convolution [34], 3D convolution [83], and 3D geometry constraint [84] outperforms other DL methods for

DE. The last factor can be the loss of functions. There is some lack of information from monocular videos, such as scale inconsistency and scale ambiguity. One of the solutions for that is using semantic information and smooth loss to learn the scales. However, increasing the loss of functions will create more complicated networks and cause more computational time.

7.2. Computational Time

Computational times depend on the number of parameters of the whole network. The complex networks can predict high-quality and accurate depths, but this will cause them to not be considered in real-time applications due to the increased consumption power requirement. One of the best ways to reduce the computational time is to use pretrained models such as ResNet [142] or DenseNet [143] for feature extractions, and the model can focus only on the decoder part of the network. Table 7 represents the comparison of complex and lightweight models developed so far for monocular depth estimation based on the NYUDv2 dataset. As shown in Table 7, there is a kind of trade-off between the accuracy and the complexity of the models. The complex models [69] (e.g., [120] with 77 million parameters (params) and 186 G floating-point operations per second (FLOPs)) require higher computational time and with a large number of trained parameters; however, they give a more accurate depth estimation. On the contrary, lightweight models (e.g., [120] with 1.7 million params and 1.5 G FLOPs) require low computational time with a low number of trained parameters. Still, the accuracy is lower than complex models. In addition, the resolution of the resulted depth images is an essential key for increasing or decreasing the computational resources for the developed MDE models.

Table 7. Comparison of complex and lightweight models based on the NYUDv2 dataset.

Group	Method	Resolution	FLOPs	Params	REL	RMS	RMSE-Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Complex	Hu et al. [144]	228 × 304	107G	67M	0.130	0.505	0.057	0.831	0.965	0.991
	Chen et al. [145]	228 × 304	150G	258M	0.111	0.420	0.048	0.878	0.976	0.993
	Yin et al. [87]	384 × 384	184G	90M	0.105	0.406	0.046	0.881	0.976	0.993
	Lee et al. [72]	416 × 544	132G	66M	0.113	0.407	0.049	0.871	0.977	0.995
	Bhat et al. [69]	426 × 560	186G	77M	0.103	0.364	0.044	0.902	0.983	0.997
Light Weight	Wofk et al. [132]	224 × 224	0.75G	3.9M	0.162	0.591	-	0.778	0.942	0.987
	Nekrasov et al. [146]	480 × 640	6.49G	2.99M	0.149	0.565	-	0.790	0.955	0.990
	Yin et al. [87]	338 × 338	15.6G	2.7M	0.135	-	0.060	0.813	0.958	0.991
	Hu et al. [147]	228 × 304	14G	1.7M	0.138	0.499	0.059	0.818	0.960	0.990
	Sheng et al. [120]	228 × 304	1.5G	8.2M	0.135	0.488	0.057	0.831	0.966	0.991

7.3. Resolution Quality

Computing a high-resolution DE is one of the main challenging tasks for researchers. Most of the current DE methods are suffering from this, and their results are not satisfied in reality. Based on the discussed training manners, it is evident that SL models [9,37,38] achieved higher quality resolution of depth maps than other models, such as UL [16,39,40]. SSL [26,41,42], because training the models with original ground truth helps the model to learn more accurately with higher quality resolution. However, one of the solutions for improving the resolution quality is to use super-resolution color images for training. However, this requires creating a new dataset which is expensive and time-consuming. In addition, the processing of high-resolution images/videos needs high computational resources that increase the cost, and obtaining high-resolution depth maps and computational resources is a trade-off.

7.4. Real-Time Inference

For using the MDE methods in industrial applications, it is very important that the model can perform in real time. There is a negative correlation between real-time

performance and the complexity of the network, as shown in Table 7. Therefore, for better performance in real-time applications, lightweight MDE networks are required. However, researchers need to consider that lightweight networks sometimes reduce the accuracy and resolution of the predicted depth maps.

7.5. Transferability

Some networks are limited to working on the exact scenarios or environments, making them useless for other types of datasets. The transferability will make them more useful for different scenarios, cameras, and datasets. Training and testing the methods on different datasets, using domain adoption technology and 3D geometry, will improve the transferability of the models, and that will cause them to become more valuable in real life.

7.6. Input Data Shapes

As discussed earlier in Section 5, there are three types of input data: mono-sequence [16,33,34], stereo sequence [7,26], and sequence-to-sequence [35,36]. The mono-sequence input shapes models receive a single image as an input and provide a single output. These types are most commonly used in UL models. On the contrary, stereo-based models receive left and right pairwise images as inputs (i.e., one pair of images is used as a target image for unsupervised learning) and provide a single output as depth maps. These input shapes are mainly used for UL and SL models. The last type, sequence-to-sequence, is necessary for RNN models. These types receive a series of images as an input and provide a sequence of depth maps as an output. Due to the simplicity of the resources for mono-sequence and sequence-to-sequence models, which require a single camera compared to the stereo models, which require at least a pair of cameras, it is more economical to use mono-sequence or sequence-to-sequence models. On the other hand, sequence-to-sequence models require higher computational resources to train the model than mono-sequence models, since they need to process a sequence of images. Therefore, the most suitable models regarding low cost and computational resources are mono-sequence models.

7.7. Future Study

The current DL methods [34,83,148] have achieved the best performance so far. However, there is still no unit network that can predict a depth with high accuracy and resolution using low computational resources and without needing the actual ground truth. Therefore, the future study can create lightweight networks working on limited-memory devices without reducing the quality and resolution of predicted depth. In addition, the developed models should achieve higher accuracy under UL models to remove the original ground truth from training and create a self-adaption network for 3D reconstruction. Currently, the main challenges of MDE are that most MDE approaches depend on high-resolution images and large-size DL models with a high number of trained parameters that help predict depth maps with high accuracy. However, these models cannot be worked in real-time applications because they require high computational time and resources. On the contrary, lightweight networks are more useful for real-time applications and can be executed on devices with limited resources. However, reducing the networks' complexity will significantly degrade the results' quality and accuracy. Therefore, there is still a gap and limitation in this area to be discovered and solved.

Accurate real-depth annotations are difficult to acquire, needing special and expensive devices such as a LIDAR sensor. Self-supervised DE methods try to overcome this problem by processing video or stereo sequences, which may not always be available. Therefore, for DE, the researchers need to cope with the issue of domain adaption that will help train a monocular depth estimation model using a fully-annotated source dataset and a non-annotated target dataset. Additionally, although the MDE networks can be trained on an alternative dataset to overcome the dataset scale problem, the trained models cannot generalize to the target domain due to the domain discrepancy. For instance, there is no general MDE network that can still correctly predict the depth maps from day and night

or indoor and outdoor images. In addition, most advanced MDE methods fail to predict accurate depth maps with adverse weather conditions (fogs, sunny, snow, etc.). Therefore, the future study requires a complete dataset to include day and night or indoor and outdoor images with different weather conditions.

8. Conclusions

DL techniques possess great potential to predict depth from monocular images. Implementation of depth prediction from monocular images is possible using an efficacious DL network structure and a dataset appropriate for the technique applied in learning. This paper presented a comprehensive overview of the contribution of this growing area of science in deep-learning-based MDE. Hence, the authors made an effort to review the state-of-the-art investigations on MDE from disparate aspects, including data input types, training manner and SL, UL, and SSL approaches combined with the application of different datasets and evaluation indicators. Finally, we highlight valuable opinions related to accuracy, computational time, resolution quality, real-time inference, transferability, and input data shapes, opening new horizons for future research. This paper demonstrates that the networks could train for various representation problems. In future perspectives, the architecture of DL models has to be improved to enhance the precision and reliability of the proposed networks and decline their inference time. Additionally, MDE networks have brilliant potential to be used in autonomous vehicles if high reliability is obtained. In addition, they must have the capability to output real-time depth maps.

Author Contributions: Conceptualization, A.M.; methodology, A.M.; software, A.M.; validation, A.M., H.A.R., M.S.A.; formal analysis, A.M., H.A.R., M.S.A., J.C.; investigation, A.M.; resources, A.M., H.A.R., J.C.; writing—original draft preparation, A.M.; writing—review and editing, H.A.R., M.S.A.; visualization, A.M.; supervision, D.P.; project administration, D.P.; funding acquisition, D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was possible with the support of the Secretariad Universitatsi Recercadel Departamentd Empresai Coneixement de la Generalitat de Catalunya (2020 FISDU 00405).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thankfully acknowledge the use of the University of Rovira I Virgili (URV) facility in carrying out this work.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DE	Depth Estimation
MDE	Monocular Depth Estimation
BDE	Binocular Depth Estimation
MVDE	Multi-View Depth Estimation
DL	Deep Learning
DI	Depth Information
CNN	Convolutional Neural Network
SL	Supervised Learning
UL	Unsupervised Learning
SSL	Semi-Supervised Learning
GTD	Ground Truth Depth
EDM	Estimated Depth Map
VO	Visual Odometry

References

1. Sun, X.; Xu, Z.; Meng, N.; Lam, E.Y.; So, H.K.H. Data-driven light field depth estimation using deep Convolutional Neural Networks. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 367–374. [CrossRef]
2. Lam, E.Y. Computational photography with plenoptic camera and light field capture: Tutorial. *J. Opt. Soc. Am. A* **2015**, *32*, 2021–2032. [CrossRef] [PubMed]
3. Khan, W.; Ansell, D.; Kuru, K.; Amina, M. Automated aircraft instrument reading using real time video analysis. In Proceedings of the 2016 IEEE 8th International Conference on Intelligent Systems (IS), Sofia, Bulgaria, 4–6 September 2016; pp. 416–420.
4. Khan, W.; Hussain, A.; Kuru, K.; Al-Askar, H. Pupil localisation and eye centre estimation using machine learning and computer vision. *Sensors* **2020**, *20*, 3785. [CrossRef] [PubMed]
5. Nomani, A.; Ansari, Y.; Nasirpour, M.H.; Masoumian, A.; Pour, E.S.; Valizadeh, A. PSOWNNs-CNN: A Computational Radiology for Breast Cancer Diagnosis Improvement Based on Image Processing Using Machine Learning Methods. *Comput. Intell. Neurosci.* **2022**, *2022*, 5667264. [CrossRef] [PubMed]
6. Rashwan, H.A.; Solanas, A.; Puig, D.; Martínez-Ballesté, A. Understanding trust in privacy-aware video surveillance systems. *Int. J. Inf. Secur.* **2016**, *15*, 225–234. [CrossRef]
7. Godard, C.; Aodha, O.M.; Brostow, G.J. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017
8. Liu, F.; Shen, C.; Lin, G.; Reid, I. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2024–2039. [CrossRef] [PubMed]
9. Eigen, D.; Puhrsch, C.; Fergus, R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. *Adv. Neural Inf. Process. Syst.* **2014**, *27*. [CrossRef]
10. Cociaş, T.T.; Grigorescu, S.M.; Moldoveanu, F. Multiple-superquadrics based object surface estimation for grasping in service robotics. In Proceedings of the 2012 13th International Conference on Optimization of Electrical and Electronic Equipment (OPTIM), Brasov, Romania, 24–26 May 2012; pp. 1471–1477. [CrossRef]
11. Kalia, M.; Navab, N.; Salcudean, T. A Real-Time Interactive Augmented Reality Depth Estimation Technique for Surgical Robotics. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 8291–8297. [CrossRef]
12. Suo, J.; Ji, X.; Dai, Q. An overview of computational photography. *Sci. China Inf. Sci.* **2012**, *55*, 1229–1248. [CrossRef]
13. Lukac, R. *Computational Photography: Methods and Applications*; CRC Press: Boca Raton, FL, USA, 2017.
14. Masoumian, A.; Kazemi, P.; Montazer, M.C.; Rashwan, H.A.; Valls, D.P. Using The Feedback of Dynamic Active-Pixel Vision Sensor (Davis) to Prevent Slip in Real Time. In Proceedings of the 2020 6th International Conference on Mechatronics and Robotics Engineering (ICMRE), Barcelona, Spain, 12–15 February 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 63–67.
15. Ming, Y.; Meng, X.; Fan, C.; Yu, H. Deep Learning for Monocular Depth Estimation: A Review. *Neurocomputing* **2021**, *438*, 14–33. [CrossRef]
16. Zhou, T.; Brown, M.; Snavely, N.; Lowe, D.G. Unsupervised learning of depth and ego-motion from video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1851–1858.
17. Khan, F.; Salahuddin, S.; Javidnia, H. Deep learning-based monocular depth estimation methods—A state-of-the-art review. *Sensors* **2020**, *20*, 2272. [CrossRef]
18. Tosi, F.; Aleotti, F.; Poggi, M.; Mattoccia, S. Learning monocular depth estimation infusing traditional stereo knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9799–9809.
19. Ramamonjisoa, M.; Lepetit, V. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
20. Schonberger, J.L.; Frahm, J.M. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113.
21. Javidnia, H.; Corcoran, P. Accurate depth map estimation from small motions. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2453–2461.
22. Scharstein, D.; Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [CrossRef]
23. Heikkila, J.; Silvén, O. A four-step camera calibration procedure with implicit image correction. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 17–19 June 1997 IEEE: Piscataway, NJ, USA, 1997; pp. 1106–1112.
24. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [CrossRef]
25. Javidnia, H.; Corcoran, P. A depth map post-processing approach based on adaptive random walk with restart. *IEEE Access* **2016**, *4*, 5509–5519. [CrossRef]
26. Kuznietsov, Y.; Stuckler, J.; Leibe, B. Semi-supervised deep learning for monocular depth map prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6647–6655.
27. Bazrafkan, S.; Javidnia, H.; Lemley, J.; Corcoran, P. Semiparallel deep neural network hybrid architecture: First application on depth from monocular camera. *J. Electron. Imaging* **2018**, *27*, 043041. [CrossRef]

28. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2002–2011.
29. Allison, R.S.; Gillam, B.J.; Vecellio, E. Binocular depth discrimination and estimation beyond interaction space. *J. Vis.* **2009**, *9*, 10. [CrossRef]
30. Palmisano, S.; Gillam, B.; Govan, D.G.; Allison, R.S.; Harris, J.M. Stereoscopic perception of real depths at large distances. *J. Vis.* **2010**, *10*, 19. [CrossRef]
31. Glennerster, A.; Rogers, B.J.; Bradshaw, M.F. Stereoscopic depth constancy depends on the subject's task. *Vis. Res.* **1996**, *36*, 3441–3456. [CrossRef]
32. Süvari, C.B. Semi-Supervised Iterative Teacher-Student Learning for Monocular Depth Estimation. Master's Thesis, Middle East Technical University, Ankara, Turkey, 2021.
33. Mahjourian, R.; Wicke, M.; Angelova, A. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5667–5675.
34. Masoumian, A.; Rashwan, H.A.; Abdulwahab, S.; Cristiano, J.; Puig, D. GCNDepth: Self-supervised Monocular Depth Estimation based on Graph Convolutional Network. *arXiv* **2021**, arXiv:2112.06782.
35. CS Kumar, A.; Bhandarkar, S.M.; Prasad, M. Depthnet: A recurrent neural network architecture for monocular depth prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 283–291.
36. Mancini, M.; Costante, G.; Valigi, P.; Ciarfuglia, T.A.; Delmerico, J.; Scaramuzza, D. Toward domain independence for learning-based monocular depth estimation. *IEEE Robot. Autom. Lett.* **2017**, *2*, 1778–1785. [CrossRef]
37. Qi, X.; Liao, R.; Liu, Z.; Urtasun, R.; Jia, J. Geonet: Geometric neural network for joint depth and surface normal estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 283–291.
38. Ummenhofer, B.; Zhou, H.; Uhrig, J.; Mayer, N.; Ilg, E.; Dosovitskiy, A.; Brox, T. Demon: Depth and motion network for learning monocular stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5038–5047.
39. Zhan, H.; Garg, R.; Weerasekera, C.S.; Li, K.; Agarwal, H.; Reid, I. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 340–349.
40. Garg, R.; Bg, V.K.; Carneiro, G.; Reid, I. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 740–756.
41. Luo, Y.; Ren, J.; Lin, M.; Pang, J.; Sun, W.; Li, H.; Lin, L. Single view stereo matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 155–163.
42. Xie, J.; Girshick, R.; Farhadi, A. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 842–857.
43. Zhao, C.; Sun, Q.; Zhang, C.; Tang, Y.; Qian, F. Monocular depth estimation based on deep learning: An overview. *Sci. China Technol. Sci.* **2020**, 63, 1612–1627. [CrossRef]
44. Dong, X.; Garratt, M.A.; Anavatti, S.G.; Abbass, H.A. Towards real-time monocular depth estimation for robotics: A survey. *arXiv* **2021**, arXiv:2111.08600.
45. Vyas, P.; Saxena, C.; Badapanda, A.; Goswami, A. Outdoor Monocular Depth Estimation: A Research Review. *arXiv* **2022**, arXiv:2205.01399.
46. Trouvé, P.; Champagnat, F.; Le Besnerais, G.; Sabater, J.; Avignon, T.; Idier, J. Passive depth estimation using chromatic aberration and a depth from defocus approach. *Appl. Opt.* **2013**, *52*, 7152–7164. [CrossRef] [PubMed]
47. Rodrigues, R.T.; Miraldo, P.; Dimarogonas, D.V.; Aguiar, A.P. Active depth estimation: Stability analysis and its applications. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 2002–2008.
48. Ulrich, L.; Vezzetti, E.; Moos, S.; Marcolin, F. Analysis of RGB-D camera technologies for supporting different facial usage scenarios. *Multimed. Tools Appl.* **2020**, *79*, 29375–29398. [CrossRef]
49. Kim, H.M.; Kim, M.S.; Lee, G.J.; Jang, H.J.; Song, Y.M. Miniaturized 3D depth sensing-based smartphone light field camera. *Sensors* **2020**, *20*, 2129. [CrossRef]
50. Boykov, Y.; Veksler, O.; Zabih, R. A variable window approach to early vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1283–1294. [CrossRef]
51. Meng, Z.; Kong, X.; Meng, L.; Tomiyama, H. Stereo Vision-Based Depth Estimation. In *Advances in Artificial Intelligence and Data Engineering*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 1209–1216.
52. Sanz, P.R.; Mezcua, B.R.; Pena, J.M.S. *Depth Estimation—An Introduction*; IntechOpen: London, UK, 2012.

53. Loop, C.; Zhang, Z. Computing rectifying homographies for stereo vision. In Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Fort Collins, CO, USA, 23–25 June 1999; IEEE: Piscataway, NJ, USA, 1999; Volume 1, pp. 125–131.
54. Fusiello, A.; Trucco, E.; Verri, A. Rectification with unconstrained stereo geometry. In Proceedings of the British Machine Vision Conference (BMVC), Colchester, UK, 8–11 September 1997; pp. 400–409.
55. Kat, R.; Jevnisek, R.; Avidan, S. Matching pixels using co-occurrence statistics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1751–1759.
56. Zhong, F.; Quan, C. Stereo-rectification and homography-transform-based stereo matching methods for stereo digital image correlation. *Measurement* **2021**, *173*, 108635. [CrossRef]
57. Zhou, K.; Meng, X.; Cheng, B. Review of stereo matching algorithms based on deep learning. *Comput. Intell. Neurosci.* **2020**. [CrossRef]
58. Alagoz, B.B. Obtaining depth maps from color images by region based stereo matching algorithms. *arXiv* **2008**, arXiv:0812.1340.
59. Luo, W.; Schwing, A.G.; Urtasun, R. Efficient deep learning for stereo matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5695–5703.
60. Aboali, M.; Abd Manap, N.; Darsono, A.M.; Mohd Yusof, Z. A Multistage Hybrid Median Filter Design of Stereo Matching Algorithms on Image Processing. *J. Telecommun. Electron. Comput. Eng. (JTEC)* **2018**, *10*, 133–141.
61. Hyun, J.; Kim, Y.; Kim, J.; Moon, B. Hardware-friendly architecture for a pseudo 2D weighted median filter based on sparse-window approach. *Multimed. Tools Appl.* **2020**, *80*, 34221–34236. [CrossRef]
62. da Silva Vieira, G.; Soares, F.A.A.; Laureano, G.T.; Parreira, R.T.; Ferreira, J.C.; Salvini, R. Disparity Map Adjustment: A Post-Processing Technique. In Proceedings of the 2018 IEEE Symposium on Computers and Communications (ISCC), Natal, Brazil, 25–28 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 00580–00585.
63. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 3354–3361.
64. Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4040–4048.
65. Zhao, C.; Tang, Y.; Sun, Q.; Vasilakos, A.V. Deep direct visual odometry. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 7733–7742. [CrossRef]
66. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding convolution for semantic segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1451–1460.
67. Xue, F.; Wang, X.; Li, S.; Wang, Q.; Wang, J.; Zha, H. Beyond tracking: Selecting memory and refining poses for deep visual odometry. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8575–8583.
68. Clark, R.; Wang, S.; Wen, H.; Markham, A.; Trigoni, N. Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
69. Bhat, S.F.; Alhashim, I.; Wonka, P. Adabins: Depth estimation using adaptive bins. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4009–4018.
70. Wang, R.; Pizer, S.M.; Frahm, J.M. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5555–5564.
71. Patil, V.; Van Gansbeke, W.; Dai, D.; Van Gool, L. Don't forget the past: Recurrent depth estimation from monocular video. *IEEE Robot. Autom. Lett.* **2020**, *5*, 6813–6820. [CrossRef]
72. Lee, J.H.; Han, M.K.; Ko, D.W.; Suh, I.H. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv* **2019**, arXiv:1907.10326.
73. Kuznietsov, Y.; Proesmans, M.; Van Gool, L. Comoda: Continuous monocular depth adaptation using past experiences. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 2907–2917.
74. Ramirez, P.Z.; Poggi, M.; Tosi, F.; Mattoccia, S.; Di Stefano, L. Geometry meets semantics for semi-supervised monocular depth estimation. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 298–313.
75. Aleotti, F.; Tosi, F.; Poggi, M.; Mattoccia, S. Generative adversarial networks for unsupervised monocular depth prediction. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
76. Pilzer, A.; Xu, D.; Puscas, M.; Ricci, E.; Sebe, N. Unsupervised adversarial depth estimation using cycled generative networks. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 587–595.
77. Watson, J.; Firman, M.; Brostow, G.J.; Turmukhambetov, D. Self-supervised monocular depth hints. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 2162–2171.

78. Yin, Z.; Shi, J. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1983–1992.
79. Casser, V.; Pirk, S.; Mahjourian, R.; Angelova, A. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January 2019; Volume 33, pp. 8001–8008.
80. Ranjan, A.; Jampani, V.; Balles, L.; Kim, K.; Sun, D.; Wulff, J.; Black, M.J. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12240–12249.
81. Gordon, A.; Li, H.; Jonschkowski, R.; Angelova, A. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8977–8986.
82. Zhou, J.; Wang, Y.; Qin, K.; Zeng, W. Unsupervised high-resolution depth learning from videos with dual networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6872–6881.
83. Godard, C.; Mac Aodha, O.; Firman, M.; Brostow, G.J. Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 3828–3838.
84. Shu, C.; Yu, K.; Duan, Z.; Yang, K. Feature-metric loss for self-supervised learning of depth and egomotion. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 572–588.
85. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 746–760.
86. Teed, Z.; Deng, J. Deepv2d: Video to depth with differentiable structure from motion. *arXiv* **2018**, arXiv:1812.04605.
87. Yin, W.; Liu, Y.; Shen, C.; Yan, Y. Enforcing geometric constraints of virtual normal for depth prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision,Seoul, Korea, 27 October–2 November 2019; pp. 5684–5693.
88. Yu, Z.; Gao, S. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1949–1958.
89. Zhao, S.; Fu, H.; Gong, M.; Tao, D. Geometry-aware symmetric domain adaptation for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9788–9798.
90. Jung, D.; Choi, J.; Lee, Y.; Kim, D.; Kim, C.; Manocha, D.; Lee, D. DnD: Dense Depth Estimation in Crowded Dynamic Indoor Scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12797–12807.
91. Alhashim, I.; Wonka, P. High quality monocular depth estimation via transfer learning. *arXiv* **2018**, arXiv:1812.11941.
92. Ma, F.; Cavalheiro, G.V.; Karaman, S. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3288–3295.
93. Guizilini, V.; Ambrus, R.; Pillai, S.; Raventos, A.; Gaidon, A. 3d packing for self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2485–2494.
94. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
95. Bian, J.; Li, Z.; Wang, N.; Zhan, H.; Shen, C.; Cheng, M.M.; Reid, I. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 35–45.
96. Saxena, A.; Sun, M.; Ng, A.Y. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 824–840. [CrossRef]
97. Saxena, A.; Sun, M.; Ng, A.Y. Make3D: Depth Perception from a Single Still Image. *AAAI* **2008**, *3*, 1571–1576.
98. Karsch, K.; Liu, C.; Kang, S.B. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2144–2158. [CrossRef]
99. Liu, M.; Salzmann, M.; He, X. Discrete-continuous depth estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 716–723.
100. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper depth prediction with fully convolutional residual networks. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 239–248.
101. Wang, C.; Buenaposada, J.M.; Zhu, R.; Lucey, S. Learning depth from monocular videos using direct methods. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2022–2030.

102. Jia, S.; Pei, X.; Yao, W.; Wong, S. Self-supervised Depth Estimation Leveraging Global Perception and Geometric Smoothness Using On-board Videos. *arXiv* **2021**, arXiv:2106.03505.
103. Vasiljevic, I.; Kolkin, N.; Zhang, S.; Luo, R.; Wang, H.; Dai, F.Z.; Daniele, A.F.; Mostajabi, M.; Basart, S.; Walter, M.R.; et al. Diode: A dense indoor and outdoor depth dataset. *arXiv* **2019**, arXiv:1908.00463.
104. Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; Westling, P. High-resolution stereo datasets with subpixel-accurate ground truth. In Proceedings of the German Conference on Pattern Recognition, Münster, Germany, 2–5 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 31–42.
105. Yang, G.; Song, X.; Huang, C.; Deng, Z.; Shi, J.; Zhou, B. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 899–908.
106. Couprie, C.; Farabet, C.; Najman, L.; LeCun, Y. Indoor semantic segmentation using depth information. *arXiv* **2013**, arXiv:1301.3572.
107. Nistér, D.; Naroditsky, O.; Bergen, J. Visual odometry. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004; IEEE: Piscataway, NJ, USA, 2004; Volume 1, p. I.
108. Goldman, M.; Hassner, T.; Avidan, S. Learn stereo, infer mono: Siamese networks for self-supervised, monocular, depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
109. Makarov, I.; Bakhanova, M.; Nikolenko, S.; Gerasimova, O. Self-supervised recurrent depth estimation with attention mechanisms. *PeerJ Comput. Sci.* **2022**, *8*, e865. [CrossRef] [PubMed]
110. Bugby, S.; Lees, J.; McKnight, W.; Dawood, N. Stereoscopic portable hybrid gamma imaging for source depth estimation. *Phys. Med. Biol.* **2021**, *66*, 045031. [CrossRef]
111. Praveen, S. Efficient depth estimation using sparse stereo-vision with other perception techniques. *Coding Theory* **2020**, 111. [CrossRef]
112. Mandelbaum, R.; Kamberova, G.; Mintz, M. Stereo depth estimation: A confidence interval approach. In Proceedings of the Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), Bombay, India, 7 January 1998; IEEE: Piscataway, NJ, USA, 1998; pp. 503–509.
113. Poggi, M.; Aleotti, F.; Tosi, F.; Mattoccia, S. Towards real-time unsupervised monocular depth estimation on cpu. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 5848–5854.
114. Cunningham, P.; Cord, M.; Delany, S.J. Supervised learning. In *Machine Learning Techniques for Multimedia*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 21–49.
115. Liu, X.; Sinha, A.; Ishii, M.; Hager, G.D.; Reiter, A.; Taylor, R.H.; Unberath, M. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE Trans. Med. Imaging* **2019**, *39*, 1438–1447. [CrossRef]
116. Abdulwahab, S.; Rashwan, H.A.; Masoumian, A.; Sharaf, N.; Puig, D. Promising Depth Map Prediction Method from a Single Image Based on Conditional Generative Adversarial Network. In Proceedings of the 23rd International Conference of the Catalan Association for Artificial Intelligence (CCIA), Tarragona, Spain, 14 October 2021. [CrossRef]
117. Li, B.; Shen, C.; Dai, Y.; Van Den Hengel, A.; He, M. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1119–1127.
118. dos Santos Rosa, N.; Guizilini, V.; Grassi, V. Sparse-to-continuous: Enhancing monocular depth estimation using occupancy maps. In Proceedings of the 2019 19th International Conference on Advanced Robotics (ICAR), Belo Horizonte, Brazil, 2–6 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 793–800.
119. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv* **2019**, arXiv:1907.01341.
120. Sheng, F.; Xue, F.; Chang, Y.; Liang, W.; Ming, A. Monocular Depth Distribution Alignment with Low Computation. *arXiv* **2022**, arXiv:2203.04538.
121. Geng, M.; Shang, S.; Ding, B.; Wang, H.; Zhang, P. Unsupervised learning-based depth estimation-aided visual slam approach. *Circuits Syst. Signal Process.* **2020**, *39*, 543–570. [CrossRef]
122. Lu, Y.; Lu, G. Deep unsupervised learning for simultaneous visual odometry and depth estimation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 2571–2575.
123. Pilzer, A.; Lathuiliere, S.; Sebe, N.; Ricci, E. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9768–9777.
124. Cho, J.; Min, D.; Kim, Y.; Sohn, K. A large RGB-D dataset for semi-supervised monocular depth estimation. *arXiv* **2019**, arXiv:1904.10230.
125. Hoiem, D.; Efros, A.A.; Hebert, M. Automatic photo pop-up. In *ACM Digital Library SIGGRAPH 2005 Papers*; Association for Computing Machinery: New York, NY, USA, 2005; pp. 577–584.

126. Masoumian, A.; Marei, D.G.; Abdulwahab, S.; Cristiano, J.; Puig, D.; Rashwan, H.A. Absolute distance prediction based on deep learning object detection and monocular depth estimation models. In Proceedings of the 23rd International Conference of the Catalan Association for Artificial Intelligence (CCIA), Tarragona, Spain, 14 October 2021. [CrossRef]
127. Dijk, T.v.; Croon, G.d. How do neural networks see depth in single images? In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 2183–2191.
128. Mousavian, A.; Pirsiavash, H.; Košecká, J. Joint semantic segmentation and depth estimation with deep convolutional networks. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 611–619.
129. Jung, H.; Kim, Y.; Min, D.; Oh, C.; Sohn, K. Depth prediction from a single image with conditional adversarial networks. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1717–1721.
130. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; Bry, A. End-to-end learning of geometry and context for deep stereo regression. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 66–75.
131. Facil, J.M.; Ummenhofer, B.; Zhou, H.; Montesano, L.; Brox, T.; Civera, J. CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11826–11835.
132. Wofk, D.; Ma, F.; Yang, T.J.; Karaman, S.; Sze, V. Fastdepth: Fast monocular depth estimation on embedded systems. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 6101–6108.
133. Chen, W.; Fu, Z.; Yang, D.; Deng, J. Single-image depth perception in the wild. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 730–738.
134. He, L.; Chen, C.; Zhang, T.; Zhu, H.; Wan, S. Wearable depth camera: Monocular depth estimation via sparse optimization under weak supervision. *IEEE Access* **2018**, *6*, 41337–41345. [CrossRef]
135. Fei, X.; Wong, A.; Soatto, S. Geo-supervised visual depth prediction. *IEEE Robot. Autom. Lett.* **2019**, *4*, 1661–1668. [CrossRef]
136. Li, R.; Wang, S.; Long, Z.; Gu, D. Undeepvo: Monocular visual odometry through unsupervised deep learning. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 7286–7291.
137. Wu, Z.; Wu, X.; Zhang, X.; Wang, S.; Ju, L. Spatial correspondence with generative adversarial network: Learning depth from monocular videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7494–7504.
138. Wang, Y.; Wang, P.; Yang, Z.; Luo, C.; Yang, Y.; Xu, W. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8071–8081.
139. Chen, Y.; Schmid, C.; Sminchisescu, C. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7063–7072.
140. Li, S.; Xue, F.; Wang, X.; Yan, Z.; Zha, H. Sequential adversarial learning for self-supervised deep visual odometry. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 2851–2860.
141. Almalioglu, Y.; Saputra, M.R.U.; de Gusmao, P.P.; Markham, A.; Trigoni, N. Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 5474–5480.
142. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
143. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
144. Hu, J.; Ozay, M.; Zhang, Y.; Okatani, T. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1043–1051.
145. Chen, X.; Chen, X.; Zha, Z.J. Structure-aware residual pyramid network for monocular depth estimation. *arXiv* **2019**, arXiv:1907.06023.
146. Nekrasov, V.; Dharmasiri, T.; Spek, A.; Drummond, T.; Shen, C.; Reid, I. Real-time joint semantic segmentation and depth estimation using asymmetric annotations. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 7101–7107.
147. Hu, J.; Fan, C.; Jiang, H.; Guo, X.; Gao, Y.; Lu, X.; Lam, T.L. Boosting Light-Weight Depth Estimation Via Knowledge Distillation. *arXiv* **2021**, arXiv:2105.06143.
148. Zhou, H.; Greenwood, D.; Taylor, S. Self-Supervised Monocular Depth Estimation with Internal Feature Fusion. *arXiv* **2021**, arXiv:2110.09482.