Monocular Depth Estimation with Deep Learning

Subjects: Computer Science, Artificial Intelligence

Contributor: Armin Masoumian, Hatem A. Rashwan, Julián Cristiano, M. Salman Asif, Domenec Puig

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 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.

Keywords: monocular depth estimation ; single image depth estimation ; deep learning ; multi-task learning

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][4][5][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][8][9]. DE has great potential for use in disparate applications, including grasping in robotics, robot-assisted surgery, computer graphics, and computational photography [10][11][12][13][14][15].

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][18][19]. DE can be functionally classified into three divisions, including monocular depth estimation (MDE), binocular depth estimation (BDE), or multi-view depth estimation (MVDE).

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][27][28].

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 [29][30]. RGB-D camera is a specific type of depth-sensing device that combines an RGB image and its corresponding depth image [31]. RGB-D cameras can be used in various devices such as smartphones and unmanned aerial systems due to their low cost and power consumption [32]. 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 [33].

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 [34][35][36]. 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 [37][38].

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 [39][40][41][42].

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.

3. Input Data Shapes for MDE Applying Deep Learning

3.1. Mono-Sequence

Monocular sequence input is mainly used for training the UL models. **Figure 1** 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 [43] 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.

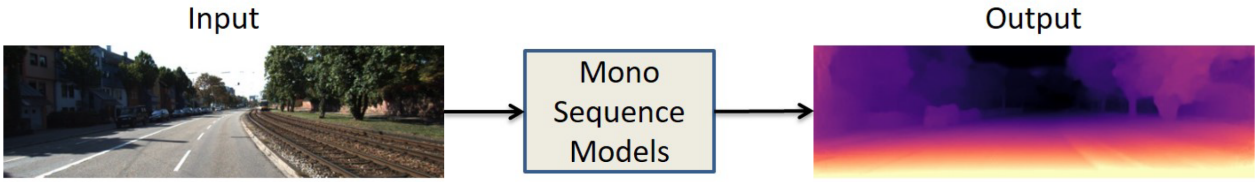

Figure 1. Data input/output structure of mono-sequence models. Single image input and single image output.

Furthermore, Mahjourian et al. [44] 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. [45] 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. [46] 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.

3.2. 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 2** 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.

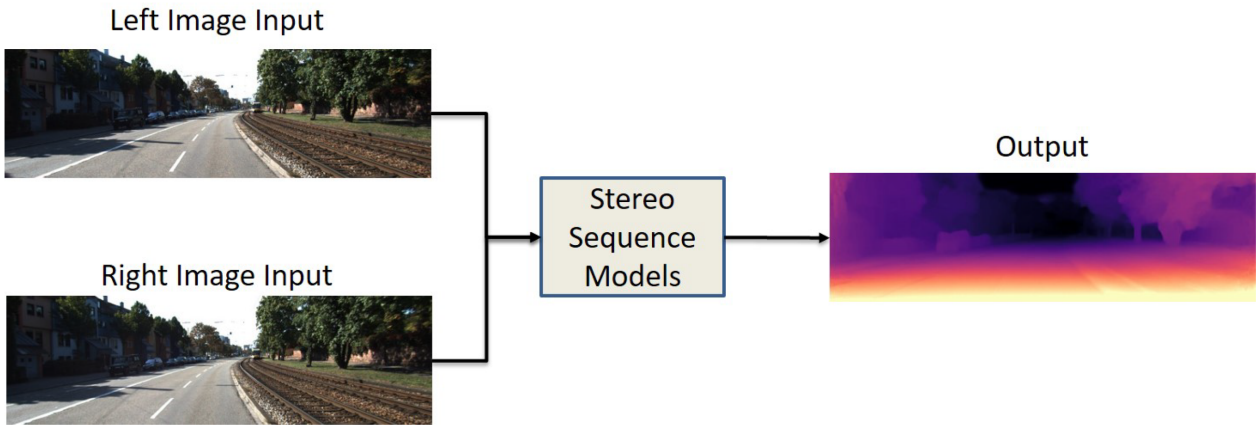

Figure 2. Data input/output structure of stereo sequence models. Stereo pairs of images as an input and single image output.

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. [47] 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.

3.3. Sequence-to-Sequence

Sequence-to-sequence data input is necessary for recurrent neural network (RNN) models [48]. These models have memory capability, which helps the system learn a group of features in sequence images. **Figure 3** 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 [48]. 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. [49] proposed an MDE model with ConvLSTM layers for learning the smooth temporal variation. Their model consists of encoder–decoder architecture. Furthermore, Mancini et al. [50] improved LSTM layers to obtain the best outcome of the predicted depth maps by feeding the input images sequentially to the system.

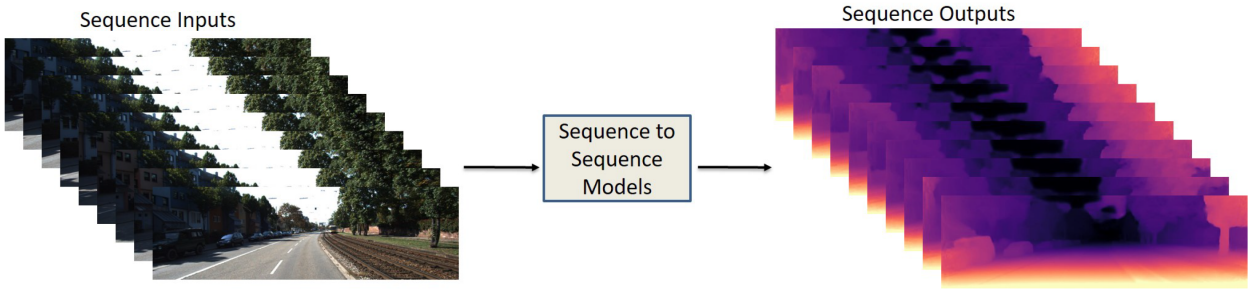

Figure 3. Data input/output structure of sequence-to-sequence models. Sequence of images as an input and sequence of images as an output.

4. 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 [51][52][53]. 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][54][55][56].

4.1. Supervised Learning Approach

The SL approach for DE needs pixel-wise ground truth DI [57]. The SL procedure applies ground truth depth (GTD) to train a neural network as a regression model [58][59][60]. 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.

After Eigen's investigation, different procedures were implemented to increase the precision of the estimated depth map (EDP). For example, Li et al. [61] 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. [54] utilized two networks to estimate the depth map and surface normal from single images.

The dataset's quality is an introductory section in SL systems, similar to methodology. Dos Santos et al. [62] 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. [63] 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. [64] 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.

4.2. 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 [65][66][67]. Garg et al. [55] 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.

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. [58] 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 [68]. Masoumian et al. [45] 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. 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.

4.3. 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.

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. [69] 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 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.

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 [70]. 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%[70].

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.
2. Lam, E.Y. Computational photography with plenoptic camera and light field capture: Tutorial. J. Opt. Soc. Am. A 2015, 32, 2021–2032.
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.
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.
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.
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.
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.
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.
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.

12. Suo, J.; Ji, X.; Dai, Q. An overview of computational photography. Sci. China Inf. Sci. 2012, 55, 1229–1248.
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.
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.
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.
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.
25. Javidnia, H.; Corcoran, P. A depth map post-processing approach based on adaptive random walk with restart. IEEE Access 2016, 4, 5509–5519.
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.
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. 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.
30. 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.
31. 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.
32. 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.
33. 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.

34. Boykov, Y.; Veksler, O.; Zabih, R. A variable window approach to early vision. IEEE Trans. Pattern Anal. Mach. Intell. 1998, 20, 1283–1294.

35. 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.

36. Sanz, P.R.; Mezcua, B.R.; Pena, J.M.S. Depth Estimation—An Introduction; IntechOpen: London, UK, 2012.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

46. 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.

47. 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.

48. Makarov, I.; Bakhanova, M.; Nikolenko, S.; Gerasimova, O. Self-supervised recurrent depth estimation with attention mechanisms. PeerJ Comput. Sci. 2022, 8, e865.

49. 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.

50. 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.

51. Bugby, S.; Lees, J.; McKnight, W.; Dawood, N. Stereoscopic portable hybrid gamma imaging for source depth estimation. Phys. Med. Biol. 2021, 66, 045031.

52. Praveen, S. Efficient depth estimation using sparse stereo-vision with other perception techniques. Coding Theory 2020, 111.

53. 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.

54. 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.

55. 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.

56. 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.
57. Cunningham, P.; Cord, M.; Delany, S.J. Supervised learning. In Machine Learning Techniques for Multimedia; Springer: Berlin/Heidelberg, Germany, 2008; pp. 21–49.
58. 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.
59. 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.
60. 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.
61. 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.
62. 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.
63. 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.
64. Sheng, F.; Xue, F.; Chang, Y.; Liang, W.; Ming, A. Monocular Depth Distribution Alignment with Low Computation. arXiv 2022, arXiv:2203.04538.
65. 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.
66. 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.
67. 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.
68. 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.
69. 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.
70. Cho, J.; Min, D.; Kim, Y.; Sohn, K. A large RGB-D dataset for semi-supervised monocular depth estimation. arXiv 2019, arXiv:1904.10230.
