A Review of Monocular Depth Estimation Methods Based on Deep Learning

*Huma Farooq[1(corresponding author)], Manzoor Ahmad Chachoo[2]

[1]Department of Computer Science, University of Kashmir, Srinagar (J&K)

[2] Department of Computer Science, University of Kashmir, Srinagar (J&K)

Email: [1]{huma.dahal@yahoo.com}

[2]{manzoorchachoo@uok.edu.in}

Abstract: In applications like autonomous vehicle driving or robot maneuverability, Precise depth estimation from images is vital for understanding the scene and its reconstruction. Traditional depth estimation techniques are based on component correspondences of several viewpoints. Monocular depth estimation from a single image is a challenging task due to the inherent ambiguity in the scene's geometry. Deep neural networks have shown great promise in addressing this problem by capturing complex features from the image and providing accurate depth maps. In this paper, we review the recent advances in monocular depth estimation based on deep learning techniques. We explore the various network frameworks and training methods used to improve the accuracy of depth estimation. We also examine the limitations of current methods and discuss open challenges in this field. Our goal is to provide a comprehensive overview of the current state-of-the-art in monocular depth estimation based on deep learning and to inspire further research in this area. The paper examines a variety of learning strategies, as well as datasets for Monocular depth estimates and challenges.

Keywords: CNN, Monocular, KITTI Dataset, NYU Dataset

1. Introduction

The term "depth estimation (DE)" refers to a set of processes and calculations used to depict a scene's spatial structure. In other words, to figure out how far apart each of the scene's points are.

Understanding and reconstructing scenes from images are among the key tasks in many applications. Recent advances in DE have concentrated on performing 2D to 3D reconstruction using convolutional neural networks (CNNs). While these

ICIDSSD 2022, March 24-25, New Delhi, India

DOI 10.4108/eai.24-3-2022.2318957

techniques' performance has steadily improved, the accuracy and resolution of these approximated depth maps are still causing considerable problems.

An approach for deducing depth information from a monocular image has recently piqued people's interest. The depth information provides crucial pieces of information or clues, such as the horizontal border, the vanishing point's location, and so on, that aid in the proficient interpretation of a scene. As a result, improvements in DE are becoming essential to developing computer vision techniques such as model-based 3D reconstruction. Because it enables for the interpretation of the geometric structure of collected images, such a depth estimate is critical for frameworks for autonomous driving. Despite substantial progress in estimating depth information using stereo images and video sequences, monocular depth estimation (MDE) remains a difficult problem because of the intrinsic uncertainty introduced by the ill-posed nature [1].

1.1 DE Methods: The methods for calculating depth are as follows:

1.1.1 Geometry-based methods: The use of geometric constrictions to recover 3D structures from images is a well-established strategy for detecting depth that has been extensively explored over the last 40 years. A delegate technique for analysing three-dimensional structures from a series of two-dimensional image configurations is structure from motion. It's a popular choice for 3D reconstruction and SLAM. Precision component coordination and image sequences of high resolution are critical to the accuracy of DE. Furthermore, SFM has a monocular scale ambiguity [2]. Stereo vision matching [3] allows you to recreate a scene's three-dimensional layout by detecting it from two different viewpoints. In order to resolve the dissimilarity maps between images, the model utilizes two cameras to emulate the human visual system. As opposed to the SFM interaction, which depends on monocular sequences, scale information is used to estimate depths in the stereo vision matching process. Although they usually rely on image sequences or image pairings, geometry-based algorithms may accurately identify the depth values of sparse spots [3]. Due to the fact that a single image provides information only in two dimensions, reconstructing a detailed depth map from it poses a significant challenge since depth is a characteristic of the three-dimensional world.

1.1.2 Sensor-based methods: Sensors that measure depth, such as LIDAR and RGB-D cameras, can quickly retrieve the resultant image's depth information. Although RGB-D cameras can directly retrieve the detailed depth map of an RGB image at the pixel level, they suffer from outside sunlight sensitivity and have a limited measuring range [4]. While LIDAR is commonly used in the industry of self-driving cars to estimate depth, it can only provide a 3D map with sparse detail. Furthermore, depth sensors' high-power consumption and large bulk limit their use in less advanced mechanics, such as drones. Over the years, the usage of monocular cameras has been on the rise due to their compact size and cost-effectiveness, resulting in an increase in the practice of inferring a depth map from a single image [5].

1.1.3 Deep learning-based methods: Deep neural networks have excelled in image processing applications including object recognition and classification due to the quick advancement of deep learning. Recent advancements have also demonstrated that deep learning can create a pixel-level depth map from start to finish from a single image. CNNs, recurrent neural networks (RNNs), and other neural networks have all been demonstrated to be good at estimating monocular depth [6], [7]. The main purpose of this paper is to grasp traditional approaches to MDE in an intuitive way. In the remaining sections, the following is summarized: Section II reviews pertinent literature. The section III discusses the most commonly used datasets and evaluation indicators for MDE. A review of deep learning methods for MDE is presented in Section IV. The review summarizes the current challenges in Section V. Finally, Section VI concludes the review.

2. Related Work

This Section summarises research on DE using images.

2.1 Traditional methods: By analysing image attributes with a nonlinear diffusion system and determining the distance between all components and the evaluated ground region's top position, Chun et al**.** were able to extract the ground section of indoor scenes. Recent research has concentrated on determining the optimal depth map for a given colour image via structural similarity learning across diverse scenes Karsch et al. [8]. The approach proposed by Torralba and Oliva is based on analyzing the frequency characteristics of an image to estimate depth. Specifically, they compute the distance between two spots in the image using a probabilistic model based on statistical properties of spectral magnitudes. This approach can be useful in situations where traditional depth DE techniques based on visual cues may not work well, such as in low-light conditions or for scenes with low texture [9]. The technique presented by Choi et al. suggests using depth gradients as a means of inferring depth instead of directly selecting depth values from the training data. The approach is founded on the notion that depth is linked to the scene's gradient and can be estimated by analyzing gradient patterns in the image. To include these depth gradients in the reconstruction process, the researchers utilized a Poisson reconstruction framework [10]. Karsch et al. developed a method to compute the depth map from a color image by analyzing its spectral features, and subsequently used a transfer scheme to refine the resulting depth map [11], [12]. Furthermore, as suggested by structural similarity, Konrad et al. sought to combine three transformation outputs adaptively, namely location-depth, color-depth, and motion-depth [13]**.**

2.2 Deep learning-based methods: Following the advent of deep neural network-based image classification, depth estimation using the generative model has identified key issues.

Ibraheem et al. suggested a transfer learning-based supervised approach for estimating depth maps. This technique estimates depth maps using a CNN. Feature extraction is based on an encoder-decoder network design that utilizes pre-trained DenseNet-169 and ImageNet networks. Moreover, the acquired data is transmitted to the decoder, which uses the sampling layer to create the final depth maps. It is trained using densified depth images, which have been expanded horizontally and colour-coded by swapping green and red channels in input images. The depth maps have a resolution of 320×240 pixels, but they are likely skewed by the bilinear up-sampling layer, which doesn't show the exact depth information for each region [14].

Using a coarse-to-fine technique to apply deep neural networks, Eigen et al. planned to directly gain proficiency with the link between colour input images and depth maps. To create the initial depth map, many convolutional layers are used, which are then fed into the next convolutional network to recover precise details from the original input. Despite the fact that the final output is hazy due to continuous convolution layer pooling processes, it displays the generative model's incredible ability to estimate the depth map from a monocular image. With enhanced organisational designs, a variety of approaches have been provided [15]. Garg et al. suggested an unsupervised technique for estimating a disparity map based on stereo reconstruction loss minimization [16]. Similarly, Godard et al. assessed the depth map in an unsupervised manner. Rather than using a predetermined set of features, the neural network was trained in a holistic manner using raw pixel values as inputs. As a result, the network is able to automatically discover and learn the most important features for the specific task it was trained on [17]. Due to the fact that both the left and right pictures can be rebuilt using the generated disparity map, their consistency loss may be used to effectively reinstate the depth information in the absence of the ground truth. It's worth noting that this methodology needs just a single image for the test stage. Despite the fact that such techniques fundamentally improve the depth estimation performance, they need an extra image, which is likely not ideal in the industry field. Gan et al. recently presented a coarse-to-fine learning strategy for multiscale systems [18]. As a result, both local and global features are taken into account for constructing a coarse depth map As the refining module learns residuals and extracts features from earlier scales and vertical pooling, the sample size of the coarse depth map increases steadily. The generative model based on deep neural networks has become very effective in detecting depth from a monocular image, however most prior techniques fail to clearly show the depth border, resulting in a blurry restoration.

3. Datasets for Depth Estimation

a) To improve the accuracy of DE, researchers rely on multiple datasets that provide a variety of images and depth maps for analysis. These datasets are selected based on their ability to cover different scenarios and object types, making them ideal for training and evaluating the performance of DE algorithms. In this paper, we explore the role of different datasets in the field of DE and our goal is to provide insights into the selection and utilization of datasets for improving the accuracy of DE [19].

Following are the most common datasets that are used to observe the scenes, as indicated in Table1.

3.1 KITTI: The dataset contains a set-up of vision tasks assembled utilizing an autonomous driving platform [20]. The full benchmark consists of numerous tasks like visual odometry, optical flow, etc. The KITTI dataset comprises of the object detection dataset, comprising the monocular images and bounding boxes. This particular dataset is widely recognized as a standard benchmark and primary source for training data when it comes to unsupervised and semi-supervised MDE. It comprises 56 scenes that have been categorized into 'city', 'residential' and 'road' classifications, with 28 scenes dedicated to each of the training and testing sets. These scenes are actual images that have been curated to enable effective training and evaluation of DE models. Furthermore, each scene contains stereo image pairs of resolution 1224×368.

3.2 NYU Depth: The NYU Depth dataset is a collection of 464 indoor video sequences that were captured using RGB-D cameras, and contain data from both depth and RGB cameras in Microsoft Kinect. This dataset is considered to be the primary resource for supervised MDE. It includes 249 indoor scenes for training and 215 for testing, with a balanced relationship between the RGB images and depth maps despite the different frame rates of the depth and RGB cameras. The dataset has been curated in such a way that each depth map is associated with its closest RGB image to ensure accurate pairing of the data. The projections from the camera are used to alter the RGB and depth sets based on the geometrical association defined by the dataset. Since of the discontinuous nature of the projection, all pixels are erased during the tests because they lack a corresponding depth value.

3.3 Cityscapes: The semantic information associated with segmentation tasks is the primary focus of the Cityscapes dataset [22]. It contains five-thousand fine labelled images and twenty-thousand coarse labelled pictures. It is composed of 22,973 stereo video sequences, acquired from 50 urban societies. It is used solely for the purpose of training unsupervised DE techniques. Furthermore, Cityscapes may be used to accurately train depth networks, and pretraining can boost performance. The various existing experiments have proved the viability of this dataset.

3.4 Make3D: It is composed of monocular depth and RGB images [23]. It does not contain any pairs of stereo images or monocular sequences. The unsupervised and semi-supervised learning strategies do not use it as the training set, but it is used in supervised approaches. This dataset is generally used as testing data for unsupervised techniques used to determine evaluate various dataset’s generalisation capability.

3.5 Pandora: Comprising of 250,000 high-resolution RGB (1920x1080 pixel) and depth (512x424) images with annotations, the Pandora dataset (24) is utilized for head centre localization and estimating shoulder and head pose. The dataset provides a comprehensive resource for research in computer vision, particularly in areas that require precise head and body orientation estimation.

Table 1. Datasets Used for MDE

Dataset	Outline
KITTI	A dataset for autonomous driving research, with stereo RGB images and depth maps provided for outdoor scenes.
NYU Depth	A dataset with indoor scenes captured by a Kinect sensor, with depth maps provided at a resolution of 640x480.
Cityscapes	A dataset for urban scene understanding, with high-resolution stereo RGB images and depth maps provided for street scenes.
Make3D	A dataset with stereo images of outdoor scenes captured by a laser scanner, with ground truth depth maps provided.
Pandora	A dataset for autonomous driving research, with stereo RGB images and depth maps provided for urban and rural scenes.

b) Evaluation Indicators in Depth Estimation: To assess depth estimate performance, "Root Mean Square Error" (RMSE), "RMSE log", "Absolute Relative Difference" (Abs Rel), "Square Relative Error" (Sq Rel), and Accuracy are used most widely [19]. Specifically, they are:

- $RMSE = \sqrt{\frac{1}{|N|}} \sum_{i \epsilon N} \ || \ d_i \ - d_i^* \, ||^2 \ )$ (1)
- $RMSE \ log = \sqrt{\frac{1}{|N|}} \sum_{i \epsilon N} || \ log(d_i) \ - log(d_i^*) \ ||^2 \ )$ (2)
- $Abs \ Rel = \frac{1}{|N|} \sum_{i \epsilon N} \ \frac{|d_i - d_i^*|}{|d_i^*|}$ (3)
- $Sq \ Rel = \frac{1}{|N|} \sum_{i \epsilon N} \ \frac{||d_i - d_i^*||^2}{|d_i^*|}$ (4)
- $Accuracies$ = % of d_i s.t. max($\frac{d_i}{d_i^*}$, $\frac{d_i^*}{di}$) = $\delta < thr$ (5)

“ d_i represents the predicted depth value of pixel *i*, and d_i^* is the ground truth of depth, *N* represents the total number of real-depth pixels, and *thr* represents the threshold of depth.”

All of the strategies depicted above were tried on the NYU-v2 or KITTI datasets. Comparing pre-trained models on both datasets has the advantage of allowing you to calculate the networks' overall performance over a large number of test sets.

4. Techniques for Monocular Depth Estimation on Deep Learning

Deep learning techniques for MDE refer to a range of methods that employ artificial neural networks with multiple layers to learn features and extract depth information from a single image. These techniques leverage the ability of deep neural networks to learn complex relationships between image features and depth cues to produce accurate depth maps from 2D images.

Users can use historical global data to anticipate the depth information contained in a single image. As a result of this, earlier works have achieved a single depth calculation by consolidating certain earlier data, analogous to the relationship between some geometric features such as the sky and ground structures [19]. With its superior image processing performance, Convolutional neural networks have also established a robust capability to properly determine dense maps from monocular pictures.

Deep neural networks use supervised signals to obtain structural information for depth inference [1]. The necessity for large datasets with expensive ground truth is the primary challenge with deep learning. This section examines methods of depth estimation through monocular vision using ground truth and a range of machine learning strategies, including supervised, semi-supervised, and unsupervised approaches. For training, supervised approaches recognise a single image and its associated depth information. In this situation, trained data is required to summarise all use cases, which is challenging.

The problem of obtaining high resolution depth estimate data as seed data has been solved by a number of semi-supervised techniques [19]. Semi-supervised algorithms are designed to train on a limited amount of labelled data alongside a larger quantity of unlabelled data. One limitation of these methods is that they rely on external information, such as camera focal length and sensor data, to make accurate predictions. In contrast, a depth network that has been trained can produce depth maps from individual images, which is a significant advantage over semi-supervised and unsupervised methods that typically require monocular sequences or stereo image pairs for training. Table 2 summarizes the relevant methods, training data, and contributions. In addition, Table 3 and Table 4 contain the quantitative results of the semi-supervised and unsupervised algorithms calculated using the KITTI dataset and the NYU Depth v2 dataset, respectively, with Graph 1 and Graph 2 displaying the graphical representation of the results.

4.1. Supervised MDE: The supervisory signal of a supervised approach is obtained from depth maps, making MDE a regressive problem [1]. Deep neural networks use individual images to predict depth maps. During the training process, the network's development is monitored using the disparities between predicted and real depth maps. Depth networks calculate scene depth information by approximating the ground truth [7] [15].

4.2. Unsupervised MDE: Unsupervised methods rely on geometric constraints between frames as a means of training, as opposed to ground truth data, which can be challenging to obtain.

The unsupervised algorithms are trained on monocular picture sequences with geometric constraints based on the distance between neighbouring frames, such as “ $(p_{n-1} \sim KT_{n\rightarrow n-1}Dn(p_n)K^{-1})$,” where “p_n ” is a pixel on image “ I_n ” and “p_{n-1}” is a matching pixel of “p_n ” on image “I_{n-1}”. The camera intrinsics matrix is denoted by “K”. “$T_{n\rightarrow n-1}$ ” measures the transformation between “ I_n ” and “I_{n-1}”, while “$D_{n(}p_n)$” measures the depth at pixel “p_n”. It is possible to establish the correlation between distinct pictures (In and In1) by applying projection functions to pixels on separate images (In and In1).

4.3. Semi-supervised MDE: Due to the lack of ground truth during training, unsupervised methods lag substantially behind supervised approaches. Unsupervised approaches are also prone to scale inconsistencies and ambiguity. Semi-supervised approaches are presented to achieve improved estimation accuracy while reducing the need on costly ground truth. The scale information can also be obtained from the semi-supervised signals. The fundamental distinction between stereo picture pairs and monocular recordings is the shift in frame order between each frame (front-back images vs. left-right images). As a result, some research label stereo image-based frameworks as unsupervised [6], while others call them semi-supervised [25]. The left-right positions between images represent the supervised signals during training, and these approaches are reviewed as semi-supervised methods [26].

Inverse depth maps (disparity maps) can be produced using semi-supervised algorithms learned on stereo image pairs. To synthesise the left picture from the right image, inverse warping is used to compute the disparity map *Dis* from predicted inverse depth. The distinction between synthetic images (I_w) and genuine visuals ($I_{l)}$ when employing supervised methods, it serves as both a signal and a constraint for the training process:

$$\text{“}L_{recons} = \sum p \, ||I_{l(}p)-I_w(p)||^2 \text{”} \tag{6}$$

$$= \sum p \, ||I_l(p)-I_r(p+Dis(p))||^2 \tag{7}$$

Table 2. Summarizing MDE Using Deep Learning

Researcher	Year	Training set	Method	Main contribution

Ibraheem et al.	2018	RGB & Depth	Supervised	CNN$_S$
Eigen et al.	2014	RGB & Depth	Supervised	CNN$_S$
Garg et al.	2016	Stereo Images	Semi-Supervised	Stereo Framework
Godard et al.	2017	Stereo Images	Semi-Supervised	Left-right consistency loss
Chen et al.	2019	Stereo Images	Semi-Supervised	Stereo matching
Chen et al.	2019	Monosequences	Unsupervised	Camera intrinsic prediction
Godard et al.	2019	Monosequences	Unsupervised	Camera intrinsic prediction

The image on the right is denoted as Ir and its depth map d is calculated using the formula d = f B/D from the projected disparity map, where f is the local camera length and B is the distance between the left and right cameras.

The semi-supervised methods are more precise than unsupervised methods because they utilize scale information obtained from semi-supervised signals. However, to ensure accuracy, it is crucial to evaluate the performance of these methods on ground truth data such as pose and LIDAR data. Despite being a more affordable alternative to dense depth maps, obtaining such data can be challenging.

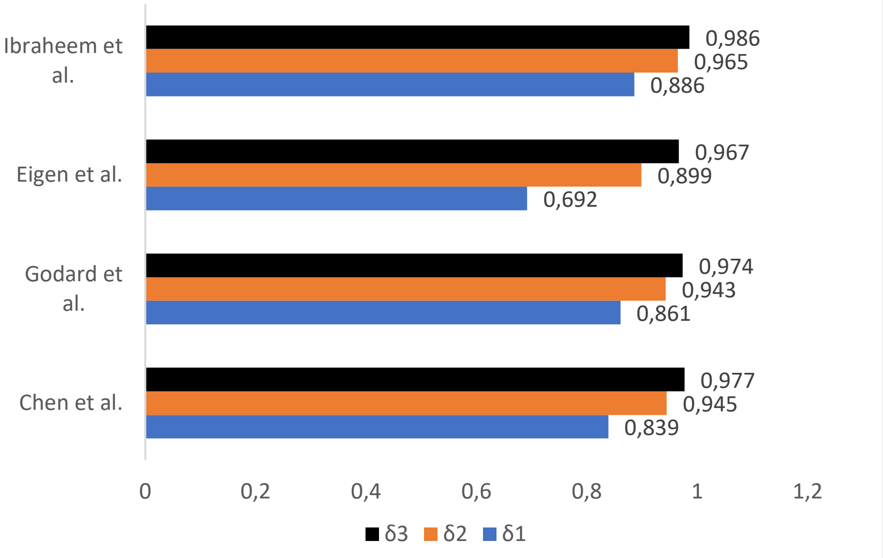
Table 3. Results of MDE on the KITTI Dataset

	(Lower value is preferred)				(Higher value is preferred)		
Method	Abs Rel	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
Chen et al.	0.118	0.905	5.096	0.211	0.839	0.945	0.977
Godard et al.	0.127	1.031	5.266	0.221	0.861	0.943	0.974
Eigen et al.	0.190	1.515	7.156	0.270	0.692	0.899	0.967
Ibraheem et al.	0.093	0.589	4.170	0.171	0.886	0.965	0.986

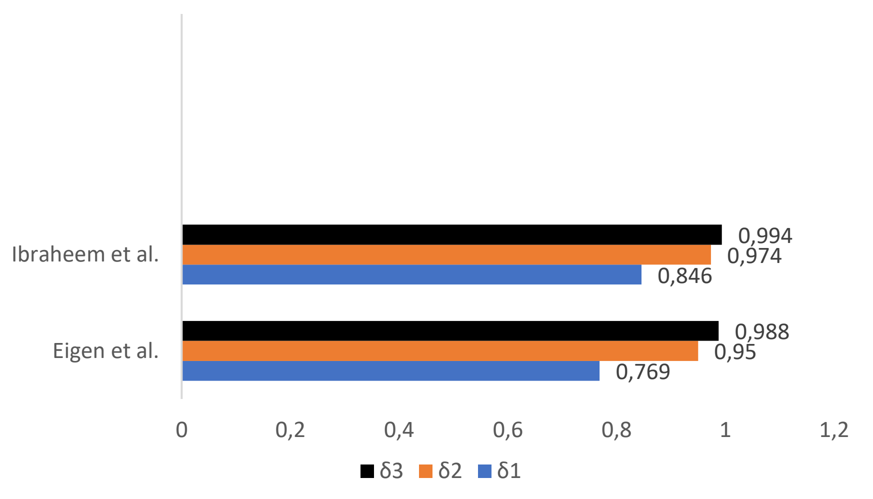
Table 4. Results of MDE on the NYU Depth v2 Dataset

Method	Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
Eigen et al.	0.158	0.641	-	0.769	0.950	0.988
Ibraheem et al.	0.123	0.465	0.053	0.846	0.974	0.994

Graph 1. Results of MDE on the KITTI Dataset



Graph 2. MDE Results on the NYU Depth v2 dataset.



5. Challenges

- *Ambiguity and Uncertainty*: The problem of DE from 2D images is challenging due to the limited information available in the 2D images, leading to inherent ambiguity in the depth perception. Even with a large dataset, it can be difficult to accurately determine the depth information from a single 2D image. Moreover, factors such as lighting conditions, occlusions, and reflections can further complicate DE, contributing to the uncertainty in the process.
- *Computational Complexity:* DE using deep learning techniques can be computationally intensive, especially for real-time applications. This requires efficient hardware and software solutions to ensure fast and reliable performance.
- *Transferability:* DE models trained on one dataset or environment may not perform well in different settings. It is essential to develop depth estimation models that can generalize well across different environments and scenarios.
- *Interpretability:* Deep learning models used in DE can be difficult to interpret, making it challenging to understand what features are used in the estimation. This can be a challenge for debugging and optimizing the model's performance.

6. Conclusion

Our primary motive is to augment research towards the MDE using deep learning techniques. The paper provides a comprehensive overview of various techniques used for MDE, covering aspects such as training methodologies and datasets utilized. The challenges are also discussed in separate section. Depth estimate is critical in MDE because of the uncertainties that can be reduced using good deep learning techniques. In image processing systems, estimating depth information is critical, hence deep learning approaches require a lot of image data.The efficient and reliable DE technique can improve the system's transferability, accuracy, real-time performance, and hence the need of an hour is the efficient DE technique.

References

1. C. Zhao, Q. Sun, C. Zhang, Y. Tang and F. Qian, "Monocular depth estimation based on deep learning: An overview", *Science China Technological Sciences*, vol. 63, no. 9, pp. 1612-1627, 2020. Available: 10.1007/s11431-020-1582-8.
2. R. Szeliski and S. B. Kang, “Shape ambiguities in structure from motion,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 5, pp. 506–512, 1997.
3. Z.-L. Cao, Z.-H. Yan, and H. Wang, “Summary of binocular stereo vision matching technology,” Journal of Chongqing University of Technology (Natural Science), vol. 29, no. 2, pp. 70–75, 2015.
4. “Cnn-slam, keisuke and tombari, federico and laina, iro and navab, nassir,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6243–6252.
5. H. Dhamo, K. Tateno, I. Laina, N. Navab and F. Tombari, "Peeking behind objects: Layered depth prediction from a single image", *Pattern Recognition Letters*, vol. 125, pp. 333-340, 2019. Available: 10.1016/j.patrec.2019.05.007.
6. R. Garg, V. K. BG, G. Carneiro, and I. Reid, “Unsupervised cnn for single view depth estimation: Geometry to the rescue,” in European Conference on Computer Vision. Springer, 2016, pp. 740–756.
7. M. Song and W. Kim, "Depth Estimation from a Single Image Using Guided Deep Network", *IEEE Access*, vol. 7, pp. 142595-142606, 2019. Available: 10.1109/access.2019.2944937.
8. C. Chun, D. Park, W. Kim, and C. Kim, “Floor detection-based depth estimation from a single indoor scene,” in Proc. IEEE Int. Conf. Image Process., pp. 3358–3362, Sep. 2013.
9. A. Torralba and A. Oliva, “Depth estimation from image structure,” IEEE Trans. Pattern Anal. Nach. Intell., vol. 24, no. 9, pp. 1226–1238, Sep. 2003.
10. S. Choi, D. Min, B. Ham, Y. Kim, C. Oh, and K. Sohn, “Depth analogy: data- driven approach for single image depth estimation using gradient samples,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 5953–5966, Dec. 2015.
11. K. Karsch, C. Liu, and S. B. Kang, “Depth transfer: Depth extraction from video using non-parametric sampling,” IEEE Trans. Pattern Anal. Nach. Intell., vol. 36, no. 11, pp. 2144–2158, Nov. 2014.

12. K. Karsch, C. Liu, and S.-B. Kang, “Depth extraction from video using non-parametric sampling,” in Proc. Eur. Conf. Comput. Vis., pp. 775–788, Oct. 2012.
13. J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, “Learning based, automatic 2D-to-3D image and video conversion,” IEEE Trans. Image Process., vol. 22, no. 9, pp. 3485–3496, Sep. 2013.
14. Alhashim, I.; Wonka, P. High quality monocular depth estimation via transfer learning. arXiv Prepr. 2018, arXiv:1812.11941.
15. D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in Proc. Neural Inform. Process. Syst., pp. 2366–2374, Dec. 2012.
16. R. Garg, G. Carneiro, and I. Reid, “Unsupervised CNN for single view depth estimation: geometry to the rescue,” in Proc. Eur. Conf. Comput. Vis., pp. 1–14, Oct. 2016.
17. C. Godard, O. M. Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., pp. 6602–6611, Jun. 2017.
18. Y. Gan, X. Xu, W. Sun, and L. Lin, “Monocular depth estimation with affinity, vertical pooling, and label enhancement,” in Proc. Eur. Conf. Comput. Vis., pp. 232–247, Sep. 2018.
19. F. Khan, S. Salahuddin and H. Javidnia, "Deep Learning-Based Monocular Depth Estimation Methods—A State-of-the-Art Review", *Sensors*, vol. 20, no. 8, p. 2272, 2020. Available: 10.3390/s20082272.
20. "kitti | TensorFlow Datasets", *TensorFlow*, 2021. [Online]. Available: https://www.tensorflow.org/datasets/catalog/kitti. [Accessed: 01- Apr- 2021].
21. "NYU Depth V2 « Nathan Silberman", *Cs.nyu.edu*, 2021. [Online]. Available: https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html. [Accessed: 18- Mar- 2021].
22. "Cityscapes Dataset – Semantic Understanding of Urban Street Scenes", *Cityscapes-dataset.com*, 2021. [Online]. Available: https://www.cityscapes-dataset.com/. [Accessed: 18- Mar- 2021].
23. "Make3D --- Range Image Dataset", *Make3d.cs.cornell.edu*, 2021. [Online]. Available: http://make3d.cs.cornell.edu/data.html. [Accessed: 18- Mar- 2021].
24. G. Borghi, "Pandora Dataset", *Aimagelab.ing.unimore.it*, 2021. [Online]. Available: https://aimagelab.ing.unimore.it/pandora/. [Accessed: 18- Mar- 2021].
25. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.
26. Farooq H, Naaz S. Performance analysis of biometric recognition system based on palmprint. International Journal of Information Technology. 2020 Dec;12(4):1281-9.