Advancements in Machine Learning-Based Depth Estimation

Hongda He

School of Science and Engineering, Chinese University of Hongkong, Shenzhen, China

121090177@link.cuhk.edu.cn

Abstract. Obtaining depth information (Depth estimation) from one or several Red-Green-Blue (RGB) images remains a crucial and challenging task in computer vision, with significant implications for applications of robotics, medical imaging, autonomous driving, augmented reality (AR), and so on. Machine learning has revolutionized depth estimation, making it the leading approach in recent years. This paper thoroughly reviews machine learning-based depth estimation methods, encompassing the classification and comparison of frequently used datasets, emerging trends, and key evaluation metrics. The study evaluates six state-of-the-art (SOTA) methods. Some common evaluation metrics are utilized to assess their performance. The results indicate that incorporating diffusion models, internal discretization, and semi-supervised learning techniques markedly enhances the performance of depth estimation models. These innovations provide valuable perspectives and experiences for future research in the field. By highlighting the effectiveness of these advanced methods, this review provides a solid foundation for understanding current capabilities and guiding future developments in machine learning-based depth estimation.

Keywords: Depth Estimation, Machine Learning, Diffusion Models, Semi-Supervised Learning.

1. Introduction

Depth Estimation aims to reconstruct the 3-dimensional (3D) geometry of the scene according to 2D images, which is a vital task in computer vision. The traditional depth estimation approaches mainly include stereo depth estimation (using two cameras to simulate the human visual system) and multi-view depth estimation. However, these approaches generally suffer from high computational complexity, occlusion, and incomplete reconstruction problems. In recent years, with advancements in deep learning, monocular depth estimation—estimating depth information according to a single Red-Green-Blue (RGB) image—has also garnered significant attention. Deep learning-based depth estimation provides a more flexible and widely applicable method. It has significant importance for applications of robotics, medical imaging, autonomous driving, augmented reality (AR), and so on.

After implementing machine learning methods, sparse depth information (e.g., sparse point clouds from Light Detection and Ranging (LiDAR)) can be completed using depth completion methods to generate dense depth maps. Ma and Karaman [1] proposed a network that learns from RGB images and sparse depth maps, combining convolutional neural networks (CNNs) and autoencoders to complete depth information. However, additional sparse-depth data sources (such as LiDAR) are still required, increasing hardware costs. Stereo vision uses images taken by two cameras to calculate depth information through disparity. For example, Pyramid Stereo Matching Network (PSMNet) enhances feature extraction with a Spatial Pyramid Pooling (SPP) [2]. This method also uses a 3D CNNs to predict disparity. Multi-View Depth Estimation is also a mainstream method. However, both methods still require complex data acquisition, increasing system complexity and cost, and require significant computational resources, which poses high real-time processing demands. Monocular depth estimation only requires a single image, has low data acquisition costs, and is widely applicable. Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth (GPLDepth) and Marigold use deep learning techniques to give each pixel an estimated value according to a single image [3, 4]. Piccinelli et al. use internal discretization techniques to improve the performance and accuracy of the deep learning method [5]. Another monocular depth estimation method is introduced by Runze Liu et al. which is based on unsupervised learning [6]. The method uses diffusion models, enhancing the model's depth learning and interpretative abilities through a hierarchical feature-guided denoising module. Depth estimation using deep learning can also be

categorized into end-to-end and non-end-to-end methods. End-to-end methods refer to the entire depth estimation process where a neural network directly generates a depth map from the input image, such as Depth Anything [7]. Non-end-to-end methods refer to the depth estimation process being divided into multiple steps, each handled separately. An initial depth map is obtained by some method and then post-processing steps (such as depth map optimization, filtering, etc.) are used to further improve the depth estimation accuracy.

This paper provides an extensive and detailed survey of depth estimation methods that are based on deep learning techniques. Its primary objective is to offer an overview of the state-of-the-art (SOTA) methods in depth estimation, detailing the principles underlying various models and comparing their distinctive characteristics. By examining recent advancements, this study aims to enhance readers' understanding of the latest developments in depth estimation, thus supporting future research.

The structure of the paper is as follows: Section 2 classifies and compares popular datasets used in the field, and outlines the core concepts and principles of different depth estimation methods. Section 3 provides an analysis and discussion of the experimental results obtained from various models. Finally, Section 4 offers a conclusion of the review and new findings, which reveal potential development directions of future research.

2. Methodology

2.1. Dataset Description and Preprocessing

To support research on image depth estimation, numerous datasets have been constructed. However, these datasets are diverse and have different characteristics. Different depth estimation methods may focus on various application scenarios, leading to differing requirements for datasets. Therefore, selecting the appropriate dataset for training and testing has become a very important task.

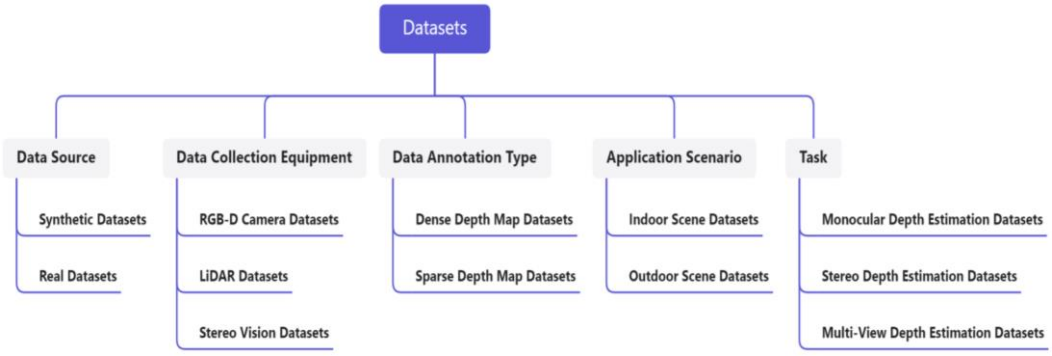

Figure 1. Taxonomy of datasets for depth estimation

This paper develops five classification criteria for existing datasets used for depth estimation, as illustrated in Fig. 1. Synthetic Datasets are generated by computers, providing accurate labels, large quantities, and the ability to cover various extreme cases [8], while Real Datasets reflect real-world scenes with higher authenticity and complexity [9-13]. Datasets may be collected by different equipment, such as RGB-D Camera [10], LiDAR [9], and Stereo Vision [12]. Data annotation types can also be different. In a dense depth map, every pixel has a depth value, which is suitable for supervised learning [10, 11]. However, in sparse depth maps, only some pixels have depth values, which are typically used in combination with other data for estimation [9]. The application scenario is also a significant classification basis, including indoor scenes and outdoor scenes. Besides, datasets can be classified according to different tasks. Table 1 shows some of the most popular datasets and compares their characteristics.

Table 1. Summary of different datasets

Datasets	Source	Annotation	Indoor	Outdoor	Images
KITTI [9]	Real	Sparse		✓	44k
Virtual KITTI [8]	Synthetic	Dense		✓	21k
NYU-v2 [10]	Real	Dense	✓		1449
Make3D [11]	Real	Dense		✓	534
CityScapes [12]	Real	Disparity		✓	5k
DIODE [13]	Real	Dense	✓	✓	25.5k

2.2. Proposed Approach

This study has corrected and accurately classified deep learning-based depth estimation methods. They are classified mainly by network architecture and degree of supervision, as shown in Fig. 2. However, the field of deep learning has experienced significant and rapid advancements, leading to numerous breakthroughs and innovations across various domains in recent years. They combine the advantages of many single architectures. In the field of depth estimation, many advanced methods are difficult to describe by a single architecture. For example, a Transformer model can be built with several encoders, decoders, and attention mechanisms [14]. Therefore, this section focuses on the SOTA methods of depth estimation in 2023 and 2024 to deeply explore the latest development trends in this field.

Figure 2. Taxonomy of deep learning-based methods

2.2.1 Diffusion for depth estimation

A diffusion model is a sophisticated type of probabilistic model employed to generate synthetic data. This model operates by simulating the process of data diffusion, where noise is gradually added to the data over a series of steps, and then, through a reverse process, the model learns to denoise the data, ultimately producing realistic and high-quality samples. These models are particularly useful in various applications such as image synthesis, text generation, and other tasks that require creating new data that closely mimics the characteristics of the original dataset. The diffusion model has been used in depth estimation [4]. However, it faces significant challenges, including high computational cost, which makes it resource-intensive to train and deploy, and sensitivity to noise, which can lead to inaccuracies and instability in the generated data.

Marigold fine-tunes the Stable Diffusion model, which has extensive visual prior knowledge [4]. However, unlike traditional diffusion models, Marigold does not perform the diffusion process directly on the original datasets. Instead, it uses a variational autoencoder to encode the depth maps and images into a low-dimensional latent space for processing. This method significantly improves computational efficiency and image generation. Marigold also modified U-Net. U-Net is an advanced network architecture based on encoder-decoder structure. The feature map from each corresponding layer in the encoder is connected directly to the decoder. These connections help restore the detailed information of the image and alleviate the problem of information loss during upsampling. U-Net is used in Marigold for the conditional denoising process. Marigold modifies the initial layer of the U-Net architecture to accept connected latent encoding, thereby enhancing the model's flexibility and adaptability to various tasks. This modification allows for improved integration of contextual

information, leading to more accurate and efficient performance across a diverse range of applications. Fig. 3 shows the architecture of Marigold.

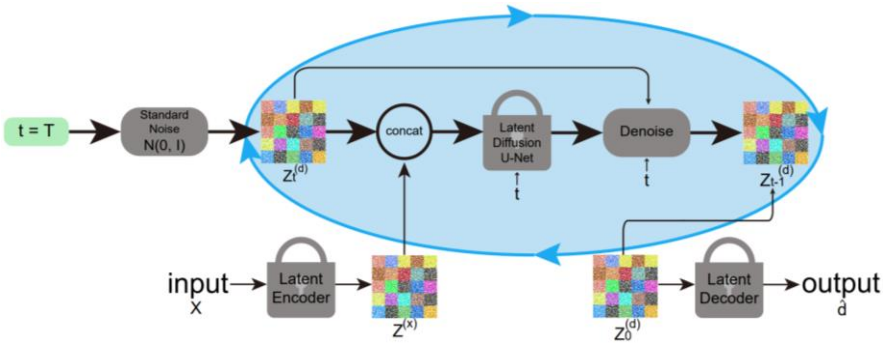

Figure 3. Overview of the Marigold inference scheme

2.2.2 Internal discretization for depth estimation

Feature discretization is the process of converting continuous features (such as numerical data) into discrete features (categorical data). This transformation maps each continuous value into a discrete category by dividing the range of values of a continuous variable into a limited number of intervals, or categories. Feature discretization can improve model performance by reducing noise and enhancing feature discrimination capabilities. It also reduces the possibility of model overfitting. Feature discretization has been applied in depth estimation. Many methods use explicit geometric constraints or convert a continuous depth range into several depth planes [5]. Although these priors improve the model's capabilities to a certain extent, they also limit the expression of the model. These models have difficulty handling the complex combination of depth patterns found in reality with various geometries or structures.

Piccinelli et al. proposed an internal discretization module. It contains a continuous-discrete-continuous bottleneck structure [5]. First, this module will perform Adaptive Feature Partitioning (AFP): After receiving feature maps of different resolutions, the model refines the initial Internal Discrete Representations (IDRs) prior through the iterative transposed cross-attention module. The iteration of each update is as follows:

$$W_{ij}^{t} = \frac{e^{k_i^T q_j^t}}{\sum_{k=1}^{N} e^{k_i^T q_k^t}} \tag{1}$$

$$q_j^{t+1} = \sum_{i=1}^{M} W_{ij}^{t} v_i \tag{2}$$

Where q_j: query, k_i: key, v_i: value, N: the number of IDRs, and M: the total number of pixels.

Given features, only a small proportion of attention weights will be much greater than zero, thus achieving soft clustering of features. Then the module will perform Internal Scene Discretization (ISD). The ISD stage combines the pixel embeddings from the decoder with the IDRs generated by the AFP stage to transfer information to a continuous output space through cross-attention layers. The internal discretization module discretizes the feature information in the feature space into a set of high-level representations. It not only does not rely on explicit geometric constraints but also improves the expressive ability and generalization of the model. It also improves information compression and processing efficiency [5].

2.2.3 Semi-supervised learning for depth estimation

Supervised learning is a machine learning approach that uses clearly labeled data for training to develop a model capable of predicting new data. Semi-supervised learning is an advanced machine learning approach that strategically combines both supervised and unsupervised learning techniques. By employing a small quantity of labeled data in conjunction with a large quantity of unlabeled data, this method aims to effectively train the model. The primary objective is to enhance model performance, leveraging the limited labeled data to guide the learning process, while simultaneously

utilizing the abundant unlabeled data to uncover underlying data structures and patterns. This approach significantly reduces the dependence on extensive labeled datasets, which are often costly and time-consuming to obtain, thereby making the training process more efficient and scalable.

Depth Anything [7] utilizes a self-training architecture: A teacher model is initially trained on labeled data to generate depth predictions and create pseudo labels for the images without depth information. The student model subsequently not only learns from the labeled data but also unlabeled data labeled by the teacher model, enhancing its ability to generalize and improve performance. This entire process is methodically divided into two distinct steps: First, the teacher model undergoes training exclusively on the labeled data for a duration of 20 epochs. This phase ensures that the teacher model can accurately predict depth information, forming a reliable basis for generating pseudo labels. secondly, the student model is trained on a mixed dataset that combines both the labeled data and the pseudo-labeled data provided by the teacher model. This mixed dataset allows the student model to benefit from the comprehensive information contained in the large-scale unlabeled images, effectively bridging the gap between labeled and unlabeled data. This approach highlights the significant value of utilizing extensive unlabeled images in depth estimation applications. By integrating both types of data, the model gains robust zero-shot capabilities, meaning it can accurately estimate depth even in scenarios where it has not been explicitly trained. This method not only enhances the overall accuracy of depth predictions but also demonstrates the potential for scalable and efficient training processes that leverage vast amounts of unlabeled data.

3. Result and Discussion

This section introduces some commonly used evaluation metrics and evaluates the previously mentioned SOTA methods using some of these metrics. The evaluation results are used to compare and study the characteristics of different methods, providing meaningful references for future research.

3.1. Evaluation Matrices

In the field of machine learning-based depth estimation, the most commonly used evaluation methods include:

$$MeanAbsoluteError(MAE) = \frac{1}{N}\sum_{i=1}^{N}\left|d_i - \hat{d}_i\right| \tag{3}$$

$$Absolute\ Re\ l\ ativeDifference(Abs\ Re\ l) = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_i - \hat{d}_i|}{d_i} \tag{4}$$

$$RootMeanSquareError(RMSE) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - \hat{d}_i)^2} \tag{5}$$

$$logMeanSquareError(logMSE) = \frac{1}{N}\sum_{i=1}^{N}(log\ d_i - \widehat{log\ d_i})^2 \tag{6}$$

$$SquareRelativeError(Sq\ Re\ l) = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_i - \hat{d}_i|^2}{d_i} \tag{7}$$

Where d_i: ground truth depth of the $i-th$ pixel, $\hat{d}_i$: predicted depth of the $i-th$ pixel, N: the total number of pixels.

Accuracy: Determine whether the ratio error between the predicted depth value and the real depth value falls within certain threshold ranges. This is typically defined as:

$$\delta_i = max(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}) \tag{8}$$

And calculate the following three metrices:

$$\delta_k = \frac{1}{N}\sum_{i=1}^{N} 1(\delta_i < 1.25^k) \tag{9}$$

Where k=1, 2, 3.

These evaluation methods can be used together to comprehensively assess the performance of depth estimation methods, helping to identify their strengths and weaknesses in different scenarios.

3.2. Result Analysis

This paper presents evaluation results for several SOTA methods on the NYU-v2 [10] dataset using metrics: accuracy, AbsRel, RMSE, and log MSE (log_{10}), as shown in Table 2. Currently, cutting-edge methods for depth estimation can already achieve an accuracy of more than 90% on the test set of NYU-v2 [10]. Compared with popular machine learning methods [3, 14, 15], the latest innovative methods [4, 5, 7] have shown better performance. Among these methods, Depth Anything [7] has the highest accuracy: δ_1, δ_2 and δ_3 are 0.984, 0.998 and 1.000 respectively. Followed by Marigold [4], δ_1 is 0.964. Depth Anything [7] also has the smallest RMSE and log_{10}, which are 0.206 and 0.024 respectively. Marigold [4] has the smallest AbsRel, which is 0.055.

Table 2. Evaluation results

Method	Higher is better			Lower is better		
	δ_1	δ_2	δ_3	AbsRel	RMSE	$log\ 1\ 0$
Adabins [14]	0.903	0.984	0.997	0.103	0.364	0.044
GPLDepth [3]	0.915	0.988	0.997	0.098	0.344	0.042
DPT [15]	0.904	0.988	0.998	0.110	0.357	0.045
Marigold [4]	0.964			**0.055**		
iDisc [5]	0.940	0.993	0.999	0.086	0.313	0.045
Depth Anything [7]	**0.984**	**0.998**	**1.000**	0.056	**0.206**	**0.024**

3.3. Discussion

The latest innovative methods have enabled machine learning-based depth estimation methods to be optimized to varying degrees. Depth Anything [7] achieved excellent results in both accuracy and error evaluations. This result indicates that unlabeled datasets have great potential to enhance the performance of depth estimation models. Using a pre-trained teacher model to generate annotations for unlabeled data can improve the generalization ability of the student model. However, this method heavily relies on the pre-trained teacher model, and if the teacher model itself performs poorly, it can constrain the performance of the final model. In the future, pre-trained models can be trained under a multi-task learning framework to do well in multiple related tasks (such as semantic segmentation), thereby providing auxiliary information for depth estimation.

Marigold and iDisc also have good evaluation results, indicating that the application of diffusion models and internal discretization operations can effectively reduce the error between predicted depth and actual depth and improve accuracy [4, 5]. However, the overall architecture of these methods remains relatively complex, presenting several challenges. Firstly, this complexity can significantly increase the difficulty of both implementation and deployment. Engineers and researchers may find it more challenging to fine-tune, debug, and maintain these sophisticated models. Additionally, complex architectures often necessitate higher computing resources, which can be a barrier to their widespread adoption, particularly in environments with limited computational capacity.

To solve these issues, future research could develop more efficient training algorithms and optimization methods. These advancements could help reduce training time and the consumption of computing resources, thereby improving overall training efficiency. For instance, techniques such as progressive learning, model pruning, and quantization might be employed to streamline the training process without compromising model performance.

Furthermore, there is a growing interest in the research and design of lightweight depth estimation models. The goal is to ensure that these models maintain high performance with a smaller number of parameters and computational complexity. This can be achieved through various approaches, such as utilizing more efficient network architectures, applying advanced compression techniques, and optimizing the model's internal representations.

By focusing on these areas, it is possible to adapt depth estimation models for resource-constrained application scenarios, such as mobile devices and real-time processing environments. Ensuring that these models are both lightweight and efficient can facilitate their deployment in a broader range of practical applications, thereby expanding their utility and impact across different fields.

4. Conclusion

This article delves into the latest advancements in machine learning-based depth estimation methods, aiming to provide researchers with a thorough understanding of both the foundational research and recent progress in this domain. The study offers a detailed classification of datasets used in depth estimation and compares the characteristics of several commonly employed datasets. It highlights recent trends in the field, including innovative applications of Diffusion Models, Internal Discretization, and Semi-Supervised Learning. Common evaluation standards are introduced and applied to assess these SOTA methods, particularly those evaluated on the NYU-v2 dataset. The findings demonstrate that these advancements significantly enhance the effectiveness of depth estimation models, offering valuable insights for future research. Nonetheless, some challenges persist, such as high model complexity and heavy reliance on pre-trained models. Future research could address these issues by investigating lightweight depth estimation models, exploring transfer learning techniques, and incorporating multi-task learning approaches. Continued innovation in depth estimation is crucial for applications across various fields, including robotics, medical imaging, autonomous driving, and AR.

References

[1] Ma F. & Karaman S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In 2018 IEEE international conference on robotics and automation (ICRA), 2018: 4796-4803.

[2] Chang J.R. & Chen Y.S. Pyramid stereo matching network. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 5410-5418.

[3] Kim D. Ka W. Ahn P. Joo D. Chun S. & Kim J. Global-local path networks for monocular depth estimation with vertical cutdepth. 2022, arXiv preprint: 2201.07436.

[4] Ke B. Obukhov A. Huang S. Metzger N. Daudt R.C. & Schindler K. Repurposing diffusion-based image generators for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 9492-9502.

[5] Piccinelli L. Sakaridis C. & Yu F. idisc: Internal discretization for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 21477-21487.

[6] Liu R. Zhu D. Zhang G. Xu Y. Shi W. Zhang X. & Li J. Unsupervised Monocular Depth Estimation Based on Hierarchical Feature-Guided Diffusion. 2024, arXiv preprint: 2406.09782.

[7] Yang L. Kang B. Huang Z. Xu X. Feng J. & Zhao H. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, 10371-10381.

[8] Gaidon A. Wang Q. Cabon Y. & Vig E. Virtual worlds as proxy for multi-object tracking analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 4340-4349.

[9] Garcia-Garcia A. Orts-Escolano S. Oprea S. Villena-Martinez V. & Garcia-Rodriguez J. A review on deep learning techniques applied to semantic segmentation. 2017, arXiv preprint: 1704.06857.

[10] Silberman N. Hoiem D. Kohli P. & Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In Computer Vision–ECCV European Conference on Computer Vision, 2012: 746-760.

[11] Puscas M. M. Xu D. Pilzer A. & Sebe N. Structured coupled generative adversarial networks for unsupervised monocular depth estimation. In 2019 International Conference on 3D Vision (3DV), 2019: 18-26.

[12] Cordts M. Omran M. Ramos S. Rehfeld T. Enzweiler M. Benenson R. & Schiele B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 3213-3223.

[13] Vasiljevic I. Kolkin N. Zhang S. Luo R. Wang H. Dai F.Z. & Shakhnarovich G. Diode: A dense indoor and outdoor depth dataset. 2019, arXiv preprint: 1908.00463.

[14] 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, 2021, 4009-4018.

[15] Ranftl R. Bochkovskiy A. & Koltun V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF international conference on computer vision, 2021: 12179-12188.