Monocular Depth Estimation Using Deep Learning: A Survey

Arya Araban
Master of Computer Science
the University of Melbourne
aaraban@student.unimelb.edu.au

Abstract: Monocular depth estimation infers dense depth maps from single RGB images, enabling 3D understanding from 2D cameras. This survey explores the evolution from early engineered methods to recent deep learning advancements. It analyzes innovations in training, network architectures, and self-supervision. Challenges include lightweight real-time networks, domain adaptation, and robustness across environments. Crucial datasets, metrics, and open problems are discussed, shaping future research on this topic.

1. Introduction

Estimating 3D structure from 2D images is a fundamental problem in computer vision with applications across robotics, augmented reality, autonomous driving, and scene understanding [1]. While LiDAR sensors capture precise geometric data, they are expensive and challenging to use widely. On the other hand, monocular cameras are cheap, widely available, and portable. Monocular depth estimation infers depth from a single RGB image, making 3D understanding possible with 2D cameras. However, due to the limited information about the 3D scene structure provided by a single view, this problem is ill-posed.

Before the advent of deep learning, monocular depth estimation relied on manually designed features and probabilistic graphical models. These approaches aimed to incorporate visual depth cues and constraints from different viewpoints to estimate depth from a single image. The advent of deep convolutional neural networks (CNNs) has led to dramatic progress by learning powerful representations directly from data. Beginning with pioneering work by Eigen et al. in 2014 [2], deep learning has become the dominant approach for monocular depth estimation.

In this review paper, we survey the field of monocular depth estimation using deep learning. We trace progress from early hand-crafted methods to modern deep network architectures. Key innovations across various training procedures are analyzed. We discuss crucial datasets, evaluation metrics, and open challenges. The goal is to provide a holistic overview of the landscape, charting the evolution of methods, models, data, and frontiers to shape future work on this important problem.

2. Traditional Depth Estimation Methods

Earlier approaches for monocular depth estimation primarily relied on hand-crafted features and graphical models. In one such approach, Saxena et al. [3] proposed a Markov Random Field (MRF) model that combined local cues and global depth cues to estimate depth from a single image. The model utilized local features such as texture gradients and haze to infer depth at a patch level, while ensuring consistency between patches through the MRF. Another approach by Hoiem et al. [4] adopted an object-based strategy, segmenting the image into geometric classes (e.g., ground, vertical, sky) and estimating depth using category-specific models. While

these methods demonstrated reasonable depth estimates within limited domains, they heavily relied on manual feature engineering and had limited learning capabilities.

More recent methods focused on enhancing generalization by training depth models from data. For instance, the Make3D approach [5] compiled a dataset consisting of images paired with corresponding laser scans to train depth estimation models. Liu et al. [6] incorporated semantic segmentations into their depth prediction process, leveraging category-specific models. However, even these approaches still depended on hand-crafted features such as textures, edges, and vanishing points. Unfortunately, these features lacked the necessary representational power to capture the complexity of real-world imagery. Consequently, these methods encountered difficulties when attempting to generalize beyond their training domains.

3. Deep Learning Depth Estimation Methods

Deep learning revolutionized the field of monocular depth estimation. By employing deep convolutional neural networks (CNNs), it became possible to learn hierarchical feature representations directly from data, streamlining the entire process. Over the past few years, these methods have made remarkable strides in performance, due to their ability to leverage large datasets and models with millions of parameters.

The input data for these networks generally fall into three main categories - mono-sequence (single image input and depth map output), stereo sequence (left and right image pair input and depth map output), and sequence-to-sequence (sequence of images as input and output depth maps). Various networks also incorporate camera pose estimation to estimate the relative camera motion between images. Each input type lends itself to different training procedures and network architectures.

The learning process in depth estimation networks can be divided into three main approaches: supervised learning, unsupervised learning, and semi-supervised learning. Each approach has its own advantages and disadvantages in terms of requiring labeled data, model complexity, and performance.

3.1 Supervised Learning

Supervised approaches rely on pixel-wise ground truth depth (GTD) maps for training. These GTD maps provide the necessary supervised signals for the neural network to learn the mapping from image pixels to depth values. Eigen et al. [2] pioneered the use of GTD maps to train CNN-based networks for monocular depth prediction. Their network consists of two deep stacks - one prepares the input image for both stacks, while the second refines the coarse depth map from the first stack to arrange the predicted depths with the scene objects. By leveraging GTD maps, which provide accurate depth information during training, they successfully trained their CNN-based network to learn the intricate relationship between image features and corresponding depth values.

To improve the accuracy of the predicted depth maps, various techniques have been developed. Li et al. [7] developed a two-stage network with conditional random fields (CRFs) for refinement. The first stage extracts image patches around superpixels (groups or regions of pixels in an image that share similar characteristics). CRFs are then applied in the second stage to refine the depth map from superpixel to pixel level. Conversely, Qi et al. [8] utilized two separate networks for depth and surface normal estimation to leverage their geometric relationship for improving accuracy. While these networks increase accuracy, they require additional ground truth like surface normals which are difficult to obtain.

The quality of training data is critical for good performance in supervised learning. Dos Santos et al. [9] generated denser GTD from sparse LIDAR measurements to enhance training data. They showed improved results compared to using just the original sparse GTD maps. Ranftl et al. [10] proposed combining multiple datasets by adapting them to a common format, avoiding overfitting to a single dataset and improving generalization. Sheng et al. [11] introduced a lightweight model utilizing an autoencoder network for depth prediction. Their innovative approach incorporated a local-global optimization scheme, integrating both local and global information to enhance the understanding of the scene's depth range and improve accuracy in depth estimation.

3.2 Unsupervised Learning

Unsupervised approaches do not require ground truth depth for training. Instead, they rely on reconstruction losses between predicted and target images to self-supervise, thus removing the need for expensive ground truth depth data. Garg et al. [12] pioneered unsupervised learning for monocular depth using stereo image pairs and reconstruction loss for self-supervision, eliminating the requirement of GTD maps. Recent advancements in unsupervised learning techniques have further expanded upon this concept, leveraging videos to estimate camera motion between consecutive frames and using it as a form of supervision.

Zhou et al. [13] simultaneously predicted depth and relative camera poses from consecutive video frames. They input three frames to a Pose CNN and Depth CNN which estimate camera motion and depth map respectively. Godard et al. [14] Further advances this method by utilizing a pose network which jointly estimates depth and camera motion from monocular videos, generating additional stereo views and frames to leverage multiple self-supervision signals within a single framework. They also address occlusion challenges through a minimum reprojection loss, achieving significant improvements over prior work.

3.3 Semi-Supervised Learning

Semi-supervised approaches aim to combine the strengths of supervised and unsupervised approaches for depth estimation and reconstruction tasks. By incorporating both ground truth depth maps and reconstruction losses, these techniques can leverage labeled and unlabeled data to generate more accurate depth predictions while minimizing errors.

Kuznietsov et al. [15] introduced a semi-supervised approach that simultaneously optimizes supervised and unsupervised losses during training. A supervised loss computed differences between estimated depth maps and ground truth, while an unsupervised reconstruction loss evaluated target views reprojected from warped source images. This multi-task framework demonstrated how leveraging both labeled and unlabeled data could enhance depth predictions.

building upon this idea of semi-supervision, Luo et al. [16] developed a two-network system to separately handle depth and view synthesis tasks. Their stereo matching network produced disparity maps from original left images and synthesized right views generated by a view synthesis network. Additionally, Luo et al. incorporated geometric constraints during inference to improve accuracy. Through this specialized multi-network design, they achieved state-of-the-art monocular depth estimation performance, outperforming the holistic model of Kuznietsov et al.

4. Datasets For Depth Estimation

A variety of datasets exist for monocular depth estimation, offering different scene types and depth ranges that cover both indoor and outdoor environments. These datasets are commonly used in deep learning methods for monocular depth estimation.

4.1 KITTI

The KITTI dataset [17] is a widely used outdoor dataset for monocular depth estimation and object detection/tracking. It was developed jointly by the Karlsruhe Institute of Technology and the Toyota Institute of Technology. The dataset includes 93,000 RGB-D training samples captured from a car equipped with high-resolution cameras, a laser scanner, and a GPS system. It covers various categories and has an image size of 1,242 × 375 with sparse ground-truth depth maps.

4.2 NYU-DEPTH V2

The NYU Depth V2 dataset [18] is an indoor dataset for monocular depth estimation using deep learning. It consists of 407,024 RGB-D image pairs captured from 464 different indoor scenes. The dataset has an original image size of 640 × 480, and the depth ranges from 0.5m to 10m. To address positional discrepancies, a coloring algorithm was used to obtain dense depth maps. The dataset includes 1,449 labeled samples for training and testing.

4.3 Make3D

The Make3D dataset [5] is an outdoor dataset for monocular depth estimation based on deep learning, and was created by Saxena et al. The dataset includes daytime city and natural scenery images, with depth maps collected using a laser scanner. It consists of 534 RGB-D image pairs, with 400 for training and 134 for testing. The depth ranges from 5m to 81m, with values beyond that uniformly mapped to 81m.

4.4 Virtual Datasets

The datasets mentioned above, namely KITTI, NYU Depth V2, and Make3D, are all derived from real-world scenes. However, there are also computer-generated virtual datasets available, such as the SceneNet RGB-D dataset [19] and the SYNTHIA dataset [20]. These virtual datasets offer a wide range of scene types with different weather, environmental, and lighting conditions.

5. Evaluation Metrics and Comparisons

The evaluation metrics proposed by Eigen et al. [2] have been widely adopted to assess and compare the performance of depth estimation methods. These metrics encompass both error and accuracy measurements. The error metrics aim for smaller values and include absolute relative error (Abs.rel), square relative error (Sq.rel), root mean square error (RMSE), and logarithm root mean square error (log RMSE). On the other hand, the accuracy rate metrics strive for higher values and involve the percentage of pixels where the predicted depth falls within a certain threshold of the ground truth depth.

- $RMSE = \sqrt{\frac{1}{T}\sum_{i \in T} \left\| d_i - d_i^{gt} \right\|^2}$
- $log\ RMSE = \sqrt{\frac{1}{T}\sum_{i \in T} \left\| log(d_i) - log(d_i^{gt}) \right\|^2}$
- $Abs.rel = \frac{1}{T}\sum_{i \in T} \frac{d_i - d_i^{gt}}{d_i^{gt}}$
- $Sq.rel = \frac{1}{T}\sum_{i \in T} \frac{\left\| d_i - d_i^{gt} \right\|^2}{d_i^{gt}}$
- $accuracies = \%\ d_i\ s.t. max\left(\frac{d_i}{d_i^{gt}}, \frac{d_i^{gt}}{d_i}\right) = \varphi < thr$

Where "d_i"and "d_i^{gt}"are the predicted and ground-truth depth respectively at the pixel indexed by "i", and "T" is the total number of pixels in all the evaluated images. "thr" represents the threshold value, which can take on the values "$(1.25)^t$" with "t" being either 1, 2, or 3.

To demonstrate how the various depth estimation approaches compare, Table 1 shows the quantitative results of the discussed supervised, unsupervised, and semi-supervised algorithms evaluated on the widely used KITTI benchmark.

Table 1 *- Evaluation Results on the KITTI dataset*

Reference	**Approach**	***Abs.rel***	***Sq.rel***	***RMSE***	***log RMSE***	$\varphi <$ **1.25**	$\varphi < (1.25)^2$	$\varphi < (1.25)^3$
		lower is better				*higher is better*		
Eigen et al. [2]	Supervised	0.215	1.515	7.156	0.270	0.692	0.899	0.967
Zhou et al. [13]	Unsupervised	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Godard et al. [14]	Unsupervised	0.114	0.898	4.935	0.206	0.861	0.949	0.976
Kuznietsov et al. [15]	Semi-supervised	0.113	0.741	4.621	0.189	0.862	0.960	0.986
Luo et al. [16]	Semi-supervised	0.094	0.626	4.252	0.177	0.891	0.965	0.984

6. Challenges and Future Work

Deep learning methods have significantly advanced performance in depth estimation. However, these models often rely on extensive computational resources and high-resolution imagery, limiting their real-time application. Additionally, the absence of a single network capable of accurately predicting depth with high resolution using low computational resources highlights an area for future research. Developing lightweight networks that operate on devices with memory constraints without compromising depth prediction quality and resolution is a pertinent challenge.

Depth estimation is further complicated by the scarcity and expense of acquiring accurate real-depth annotations, often requiring specialized equipment such as LIDAR sensors. Self-supervised depth estimation methods utilizing video or stereo sequences, are a promising alternative but are not always accessible.

Additionally, current models have difficulty generalizing to different domains. For example, we lack a

universal network capable of accurately predicting depth maps across various conditions such as day and night or indoor and outdoor environments, and in various weather conditions. Hence, future research in depth estimation should focus on creating models robust to these environmental factors. Achieving this will likely require comprehensive datasets encapsulating a wide range of conditions and environments.

7. Conclusion

In this paper, we have reviewed the evolution of monocular depth estimation, from early hand-crafted methods to dramatic advances enabled by deep learning. Beginning with pioneering CNN approaches, supervised, unsupervised, and semi-supervised frameworks have been developed to learn powerful representations straight from data. Key datasets and metrics have fueled progress, but challenges remain around lightweight real-time networks, domain adaptation, and robustness across environments and conditions. As methods continue advancing, monocular depth promises to enable ubiquitous 3D understanding from readily available 2D cameras, opening doors for applications in robotics, autonomous driving, augmented reality, and scene understanding.

References

[1] A. Saxena, S. H. Chung, and A. Y. Ng, "3-D Depth Reconstruction from a Single Still Image," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 53–69, Jan. 2008, doi: 10.1007/s11263-007-0071-y.

[2] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, in NIPS'14. Cambridge, MA, USA: MIT Press, Dec. 2014, pp. 2366–2374.

[3] A. Saxena, S. Chung, and A. Ng, "Learning Depth from Single Monocular Images," in *Advances in Neural Information Processing Systems*, Y. Weiss, B. Schölkopf, and J. Platt, Eds., MIT Press, 2005.

[4] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic Photo Pop-Up," in *ACM SIGGRAPH 2005 Papers*, in SIGGRAPH '05. New York, NY, USA: Association for Computing Machinery, 2005, pp. 577–584. doi: 10.1145/1186822.1073232.

[5] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D Scene Structure from a Single Still Image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009, doi: 10.1109/TPAMI.2008.132.

[6] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016, doi: 10.1109/TPAMI.2015.2505283.

[7] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and Surface Normal Estimation From Monocular Images Using Regression on Deep Features and Hierarchical CRFs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015.

[8] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[9] N. dos S. Rosa, V. Guizilini, and V. Grassi, "Sparse-to-Continuous: Enhancing Monocular Depth Estimation using Occupancy Maps," in *2019 19th International Conference on Advanced Robotics (ICAR)*, 2019, pp. 793–800. doi: 10.1109/ICAR46387.2019.8981652.

[10] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards Robust Monocular

Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, 2022, doi: 10.1109/TPAMI.2020.3019967.
[11] F. Sheng, F. Xue, Y. Chang, W. Liang, and A. Ming, "Monocular Depth Distribution Alignment with Low Computation," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 6548–6555. doi: 10.1109/ICRA46639.2022.9811937.
[12] R. Garg, V. K. B.G., G. Carneiro, and I. Reid, "Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 740–756.
[13] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised Learning of Depth and Ego-Motion From Video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.
[14] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging Into Self-Supervised Monocular Depth Estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.
[15] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-Supervised Deep Learning for Monocular Depth Map Prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.
[16] Y. Luo *et al.*, "Single View Stereo Matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.
[17] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 3354–3361. doi: 10.1109/CVPR.2012.6248074.
[18] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012, pp. 746–760. doi: 10.1007/978-3-642-33715-4_54.
[19] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-Training on Indoor Segmentation?," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.
[20] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.