Monocular Depth Estimation using Deep Learning: A Review

Shantanu Shukla

Master of Technology in Artificial Intelligence and Data Science
CDAC, Noida
{shantanu.shukla10@gmail.com}

Abstract

Monocular depth estimation (MDE) is a fundamental task in computer vision with applications in robotics, autonomous driving, augmented reality, and more. This review paper critically analyzes four influential deep learning-based MDE models: Monodepth, AdaBins, BinsFormer, and ManyDepth2. These models represent major shifts in learning paradigms from unsupervised stereo-based training to transformer-driven adaptive binning and motion-aware depth estimation. This paper compares their methodologies, architectural designs, performance, datasets, and application domains. It concludes by highlighting current research gaps and proposing potential future directions.

Keywords: Monocular depth estimation, deep learning, AdaBins, BinsFormer, Monodepth, ManyDepth2, computer vision

1. Introduction

Depth estimation from a single image, or monocular depth estimation (MDE), plays a pivotal role in enabling machines to perceive 3D structure from 2D inputs. Unlike stereo or LiDAR-based methods, monocular approaches are lightweight and hardware-independent. The success of deep learning has significantly enhanced MDE performance through various supervised, self-supervised, and transformer-based models. This review explores four state-of-the-art MDE models and provides a comparative analysis to understand their impact and future potential.

2. Background

Monocular depth estimation has evolved from simple convolutional networks to sophisticated frameworks utilizing attention, motion cues, and adaptive strategies. Common datasets include KITTI, NYUv2, and Make3D. Standard evaluation metrics are Absolute Relative Error (Abs Rel), Root Mean Squared Error (RMSE), and accuracy thresholds (e.g., δ1).

3. Literature Review

Recent literature in monocular depth estimation demonstrates rapid progress in both supervised and self-supervised paradigms. Godard et al. (2019) pioneered unsupervised methods by introducing left-right consistency using stereo pairs, setting the stage for methods that do not require ground-truth depth. Bhat et al. (2021) introduced AdaBins,

which utilized adaptive binning through transformer modules, achieving state-of-the-art results on indoor and outdoor benchmarks. Li et al. (2022) extended this binning concept by integrating it into a hierarchical transformer framework in BinsFormer, improving generalization and structural accuracy. Zhou et al. (2023) addressed the challenge of dynamic scenes by introducing ManyDepth2, which leveraged temporal and motion-aware features to enhance depth prediction. These papers reflect a transition from rigid convolution-based systems to flexible, attention-based, and scene-adaptive architectures, illustrating the evolving complexity and capability of MDE systems.

4. Model Reviews

4.1 Monodepth2 (2019)

- *Authors:* Godard et al.
- *Approach:* Unsupervised using stereo pairs and left-right consistency.
- *Key Components:* CNN encoder-decoder, photometric loss, left-right disparity.
- *Performance:* ~81% δ1 accuracy on KITTI.
- *Strengths:* No depth ground truth needed.
- *Weaknesses:* Limited to stereo training data.

4.2 AdaBins (2021)

- *Authors:* Bhat et al.
- *Approach:* Supervised learning with adaptive depth binning.
- *Key Components:* EfficientNet encoder, transformer-based bin regressor, bin width prediction.
- *Performance:* ~92.8% δ1 on NYUv2.
- *Strengths:* High accuracy with flexible binning.
- *Weaknesses:* Requires depth supervision.

4.3 BinsFormer (2022)

- *Authors:* Li et al.
- *Approach:* Transformer-based supervised depth estimation with bins.
- *Key Components:* Hierarchical ViT, adaptive bins, bin decoder.
- *Performance:* ~93.1% δ1 on NYUv2 and KITTI.
- *Strengths:* High precision, better generalization.
- *Weaknesses:* Complex architecture and training.

4.4 ManyDepth2 (2023)

- *Authors:* Zhou et al.

- *Approach:* Self-supervised, multi-frame, motion-aware.
- *Key Components:* Dynamic object masking, temporal feature alignment.
- *Performance:* ~90% δ1 on KITTI.
- *Strengths:* Works in dynamic scenes.
- *Weaknesses:* Higher computational load.

5. Comparative Analysis

Table 1. Comparative analysis of the performance of the four revied models.

MODEL	YEAR	TYPE	DATASET(S)	ACCURACY (Δ1)	HIGHLIGHTS
MONODEPTH2	2019	Unsupervised	KITTI	~81%	Stereo pairs, consistency loss
ADABINS	2021	Supervised	NYUv2, KITTI	~92.8%	Adaptive depth bins
BINSFORMER	2022	Supervised	NYUv2, KITTI	~93.1%	Transformer + binning
MANYDEPTH2	2023	Self-Supervised	KITTI	~90%	Dynamic scene handling

6. Strengths and Limitations

Monodepth2: Offers a lightweight, unsupervised solution well-suited for applications without access to labeled data. However, its reliance on stereo pairs for training limits its flexibility and applicability across varying scene types.

AdaBins: Delivers high accuracy using a supervised framework with adaptive binning, ideal for structured indoor scenes. Yet, it demands extensive labeled depth data, making it less suitable for unstructured or novel environments.

BinsFormer: Combines the strengths of adaptive binning with transformers for high performance and generalization. Despite its precision, its intricate architecture and computational complexity may pose challenges in real-time or resource-constrained settings.

ManyDepth2: Tailored for dynamic scenes with self-supervised, multi-frame input, offering temporal robustness. However, the model's dependency on motion cues and increased computational burden may reduce inference speed and accessibility.

7. Gaps in Research

Handling real-time performance in dynamic scenes: While ManyDepth2 addresses motion, most models struggle to maintain high accuracy with low latency in changing environments.

Cross-domain generalization without fine-tuning: Many current models perform well on specific datasets but degrade when exposed to new domains, highlighting the need for more adaptive learning techniques.

Efficient transformer models for edge devices: Transformer-based models like BinsFormer achieve excellent accuracy but often require large resources, limiting their use on mobile or embedded systems.

Combining supervision strategies effectively: There is a lack of unified frameworks that integrate supervised, self-supervised, and unsupervised learning paradigms, which could improve robustness and reduce dependency on large-scale labeled datasets.

8. Conclusion

This review traced the evolution of monocular depth estimation through four influential models, each introducing distinct innovations. While progress is clear in terms of accuracy and adaptability, challenges remain in speed, generalization, and complexity. Bridging these gaps will define the next generation of depth estimation systems.

9. References

[1] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging Into Self-Supervised Monocular Depth Estimation," in *Proc. ICCV*, 2019. [Online]. Available: https://arxiv.org/abs/1806.01260

[2] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth Estimation using Adaptive Bins," in *Proc. CVPR*, 2021. [Online]. Available: https://arxiv.org/abs/2011.14141

[3] Z. Li, X. Wang, X. Liu, and J. Jiang, "BinsFormer: Revisiting Adaptive Bins for Monocular Depth Estimation," *arXiv preprint*, 2022. [Online]. Available: https://arxiv.org/abs/2204.00987

[4] K. Zhou et al., "ManyDepth2: Motion-Aware Self-Supervised Multi-Frame Monocular Depth Estimation in Dynamic Scenes," *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2312.15268