A Comprehensive Report on Monocular Depth Estimation using Deep Learning

Submitted by

Shantanu Shukla
(01211805424)

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

Center for Development of Advanced Computing, Noida

Affiliated to Guru Gobind Singh Indraprastha University

CONTENTS

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

Monocular depth estimation (MDE) is the task of predicting a depth map from a single RGB image. It is a fundamental problem in computer vision and has important applications in robotics, autonomous driving, augmented reality, and more. Unlike stereo vision or LiDAR-based depth sensing, monocular methods do not require complex hardware setups, making them attractive for lightweight and scalable systems. The introduction of deep learning has significantly improved MDE performance, with architectures evolving from simple convolutional neural networks to sophisticated transformer-based systems. This report reviews recent advancements in MDE and analyses four state-of-the-art models: Monodepth2, AdaBins, BinsFormer, and ManyDepth2.

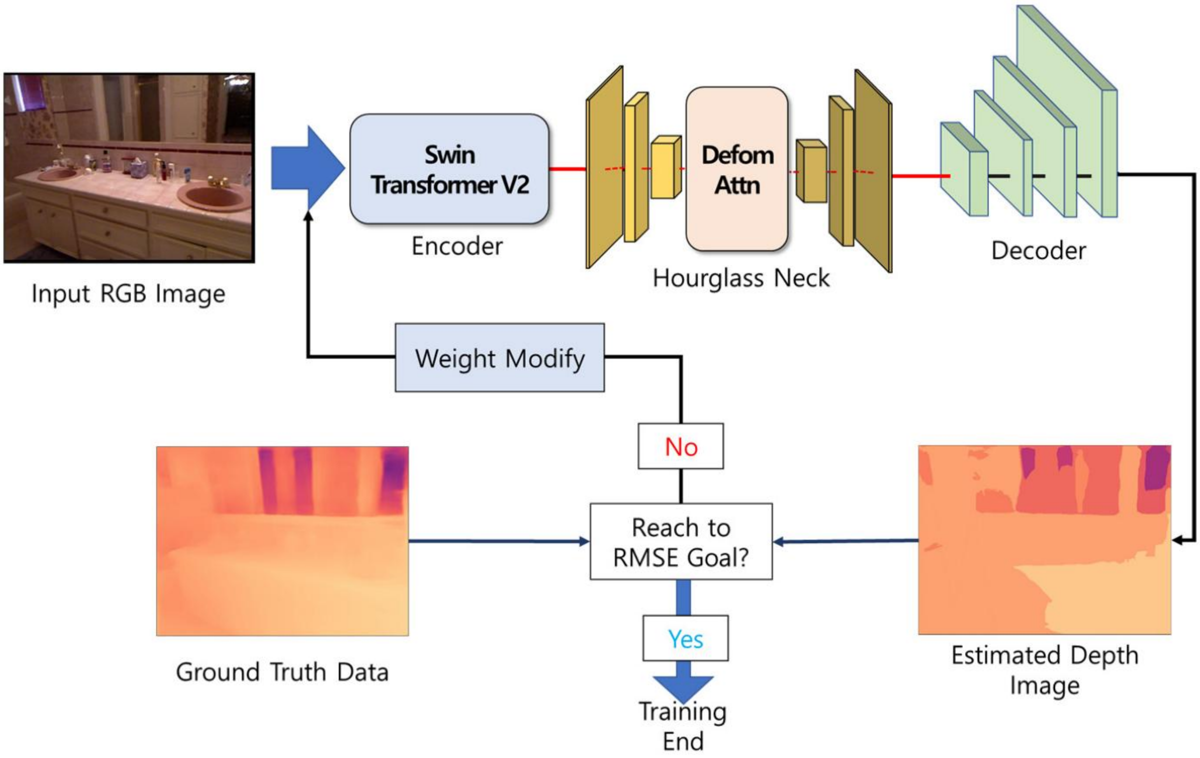

Figure 1: Overview of Monocular Depth Estimation Workflow.

CHAPTER 2: OBJECTIVE

1. **To Review Recent Progress:** Analyse the recent advancements in monocular depth estimation with a focus on deep learning-based models.
2. **Model Evaluation:** Provide an in-depth review of four recent state-of-the-art models—Monodepth2, AdaBins, BinsFormer, and ManyDepth2.
3. **Architecture Understanding:** Understand the architectural components and working principles of these models including CNNs, transformers, and encoder-decoder structures.
4. **Performance Comparison:** Compare and contrast these models based on accuracy metrics, speed, efficiency, robustness to dynamic scenes, and deployment feasibility.
5. **Highlight Challenges:** Identify the existing gaps in monocular depth estimation research and explore the limitations of current techniques.
6. **Discuss Real-World Applications:** Examine the application domains where monocular depth estimation plays a critical role.
7. **Support Future Research:** Provide a knowledge base for researchers aiming to improve MDE or apply it in new domains by outlining directions for future work.

These objectives help provide a structured approach to understanding, analysing, and critically evaluating the state of monocular depth estimation research.

CHAPTER 3: LITERATURE SURVEY

3.1 "Digging Into Self-Supervised Monocular Depth Estimation” by C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, 2019 [1]

3.1.1 Overview

Introduced by Godard et al. in 2019, Monodepth2 leverages self-supervised learning for monocular depth estimation using stereo pairs and video sequences. The model features an encoder-decoder architecture with ResNet-based encoders. It uses a photometric consistency loss to compare pixel intensities in adjacent frames, employing image reconstruction as supervision. Key innovations include per-pixel minimum reprojection loss, auto-masking for dynamic objects, and a multi-scale depth estimation strategy.

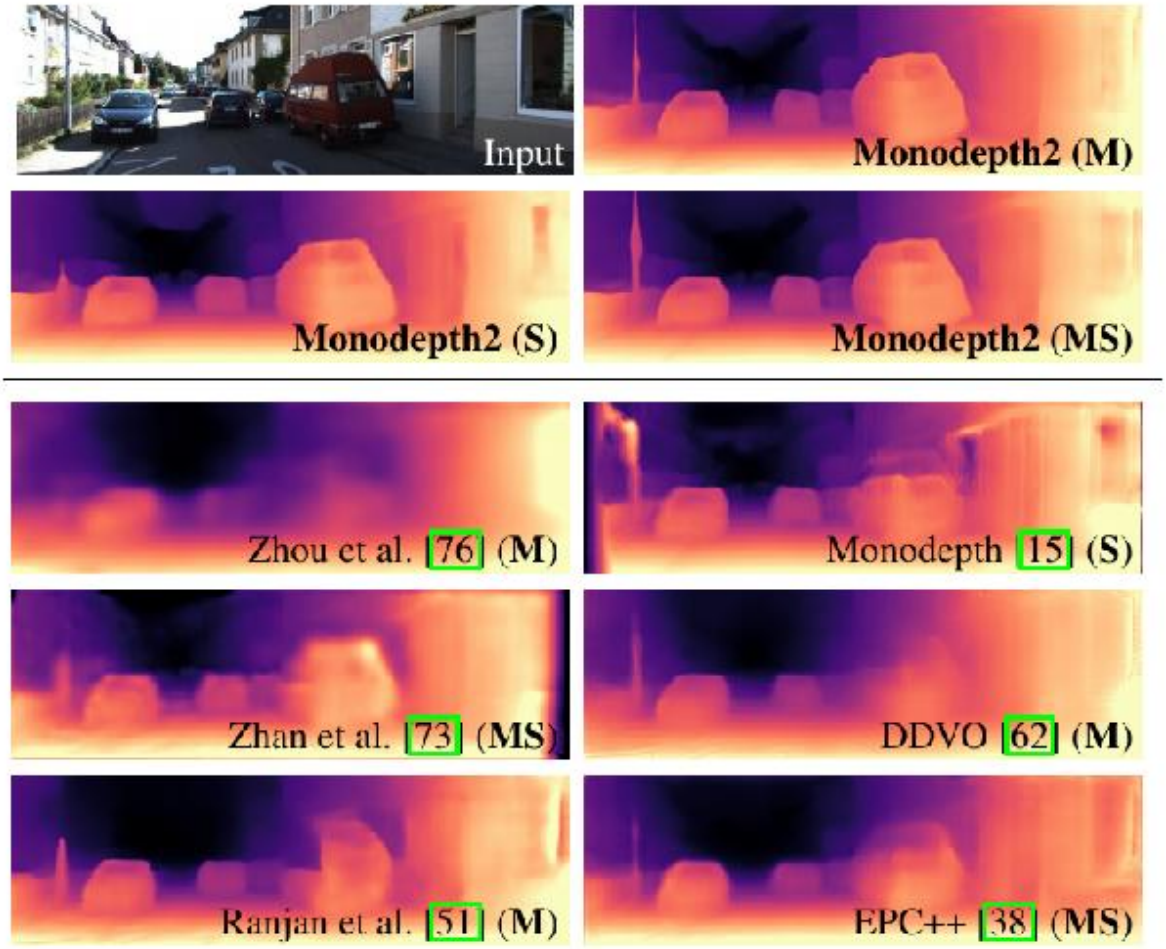

Figure 2: Depth from a single image.

3.1.2 Model Architecture

Encoder: ResNet-18 or ResNet-50.

Decoder: Convolutional upsampling layers to predict depth at multiple scales.

Loss Functions:

- Photometric loss (reconstruction between consecutive frames).
- Auto-masking to ignore dynamic object inconsistencies.
- Smoothness regularization for edge-aware predictions.

Key Features:

- Multi-scale depth prediction.
- Temporal supervision via view synthesis.

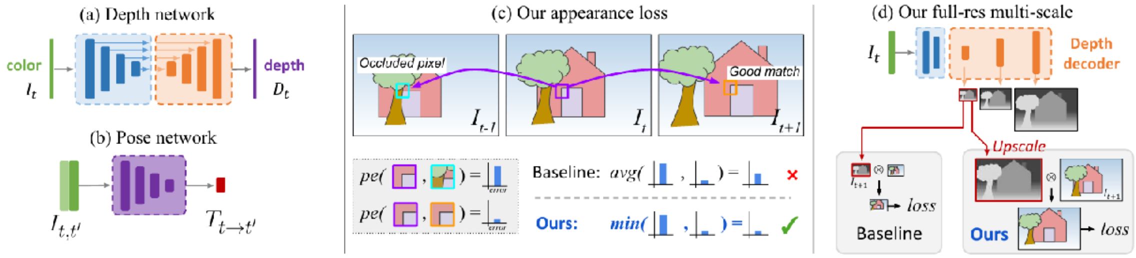

Figure 3: MonoDepth2 architecture.

3.1.3 Performance Table

Dataset	Abs Rel	Sq Rel	RMSE	RMSE log
KITTI	0.115	0.903	4.863	0.193

Table 1: MonoDepth2 performance.

3.2 “AdaBins: Depth Estimation using Adaptive Bins” by S. F. Bhat, I. Alhashim, and P. Wonka, 2021 [2]

3.2.1 Overview

Proposed by Bhat et al. in 2021, AdaBins replaces traditional continuous depth regression with a classification problem using adaptive binning. A Vision Transformer (ViT) encoder processes image features, and a decoder predicts bin centers for depth values dynamically. This approach improves resolution of depth edges and sharp boundaries.

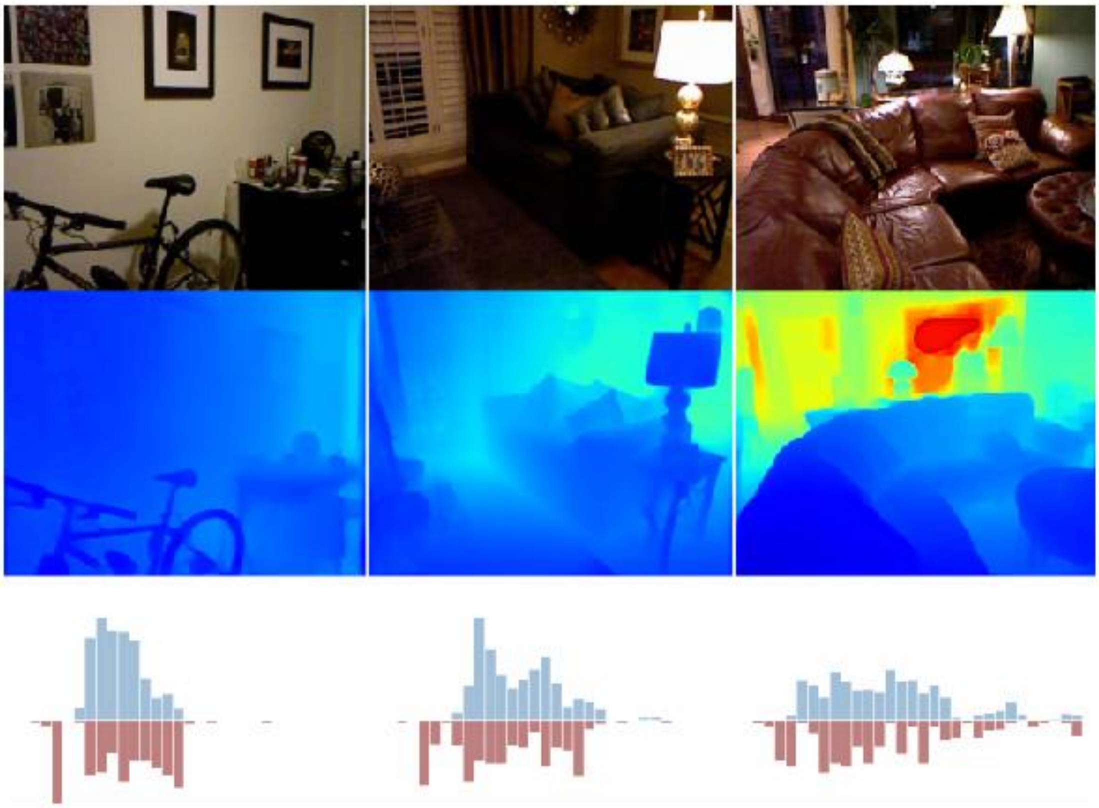

Figure 4: Illustration of AdaBins.

3.2.2 Model Architecture

Encoder: Vision Transformer (ViT-B/16 or ResNet hybrid).

Decoder:

- Bin Width Estimator: Predicts adaptive bin centers.
- Softmax Classification over bins.

Key Features:

- Converts depth regression to classification.
- Adaptive bins allow flexible depth distribution modeling.

Training:

- Cross-entropy loss on bin predictions.
- Auxiliary losses to align predicted bins with ground truth.

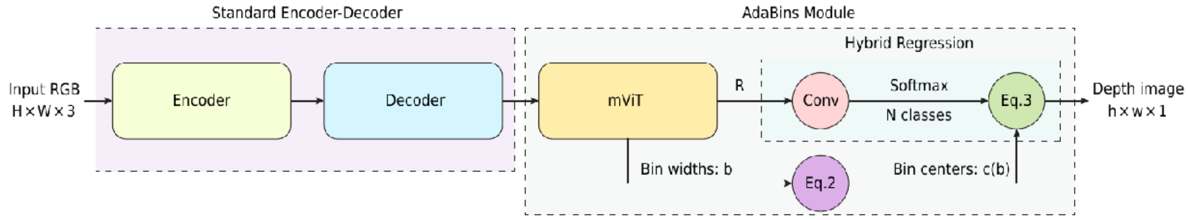

Figure 5: AdaBins architecture.

3.2.3 Performance Table

Dataset	Abs Rel	Sq Rel	RMSE	RMSE log
NYU	0.096	0.620	0.600	0.135

Table 2: AdaBins performance.

3.3 “BinsFormer: Revisiting Adaptive Bins for Monocular Depth Estimation” by Z. Li, X. Wang, X. Liu, and J. Jiang, 2022 [3]

3.3.1 Overview

BinsFormer (2022) enhances AdaBins using a Transformer-based decoder and masked attention for spatial context integration. It maintains adaptive binning but incorporates cross-attention between image and bin embeddings. The result is improved depth estimation at object boundaries and over large spatial areas.

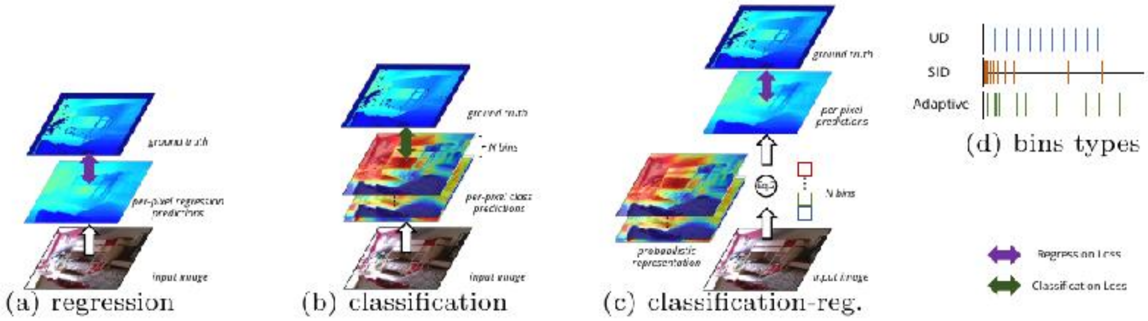

Figure 6: Regression vs. classification vs. classification-regression.

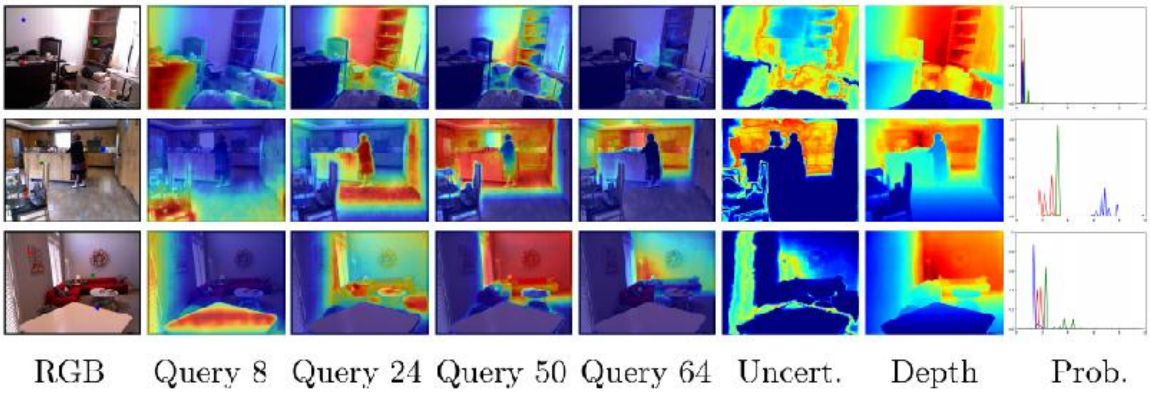

Figure 7: Visualization of depth estimation in BinsFormer.

3.3.2 Model Architecture

Encoder: Swin Transformer or ViT.

Decoder:

- Bin Decoder: Learns discrete depth bins.
- Cross-Attention Module: Integrates image features with bin embeddings.

Key Features:

- Spatially aware bin refinement.
- Robust to large and small object boundaries.
- Improved boundary sharpness over AdaBins.

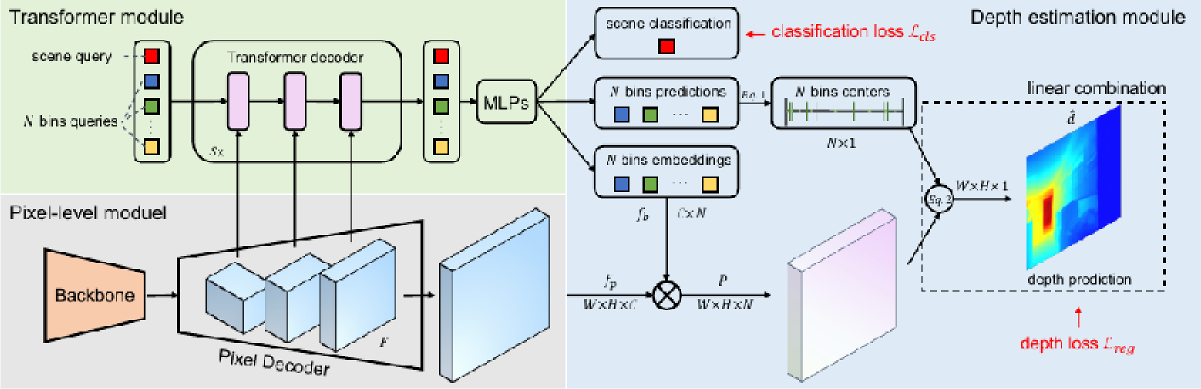

Figure 8: BinsFormer Architecture.

3.3.3 Performance Table

Dataset	Abs Rel	Sq Rel	RMSE	RMSE log
NYU	0.096	0.620	0.600	0.135

Table 3: BinsFormer performance.

3.4 “Manydepth2: Motion-Aware Self-Supervised Monocular Depth Estimation in Dynamic Scenes” by K. Zhou et al., 2023 [4]

3.4.1 Overview

ManyDepth2 (2023) focuses on motion-aware self-supervised monocular depth estimation for dynamic scenes. It combines motion segmentation, a depth CNN, and a pose network trained with photometric consistency. Optical flow is used to estimate scene motion and improve depth prediction for moving objects.

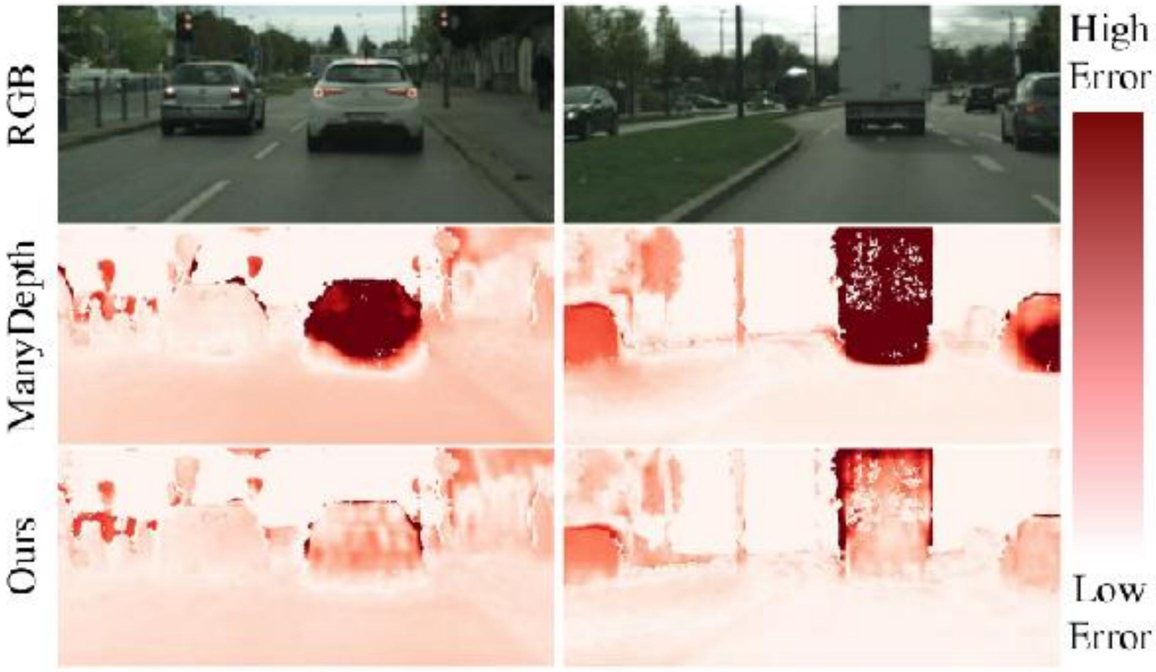

Figure 9: Qualitative comparison on Cityscapes.

3.4.2 Model Architecture

Encoder: ResNet for feature extraction.

Pose Network: Separate network to estimate ego-motion.

Depth Decoder:

- Uses learned features and temporal data.
- Integrated with motion segmentation for moving object handling.

Key Features:

- Combines motion segmentation + photometric loss.
- Trains with video data to exploit temporal consistency.
- Optimized for dynamic scenes.

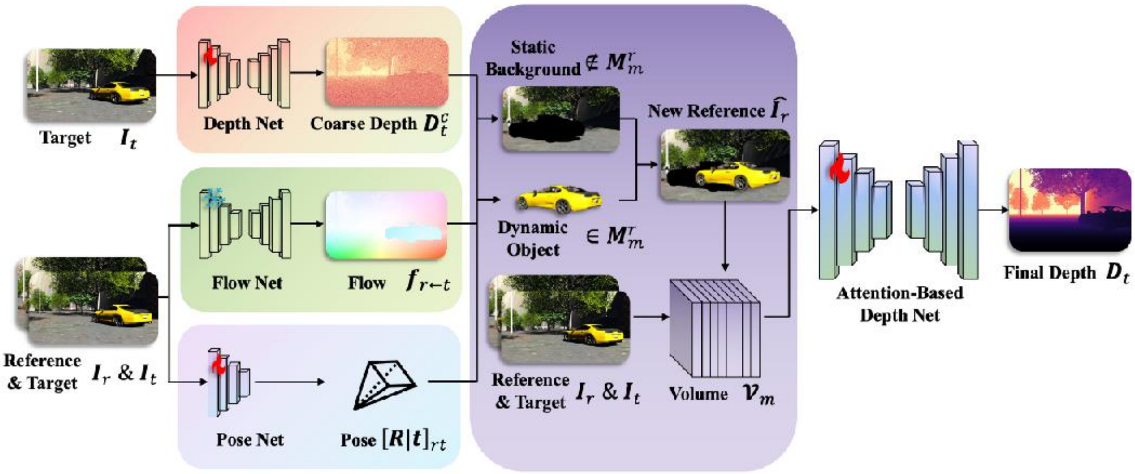
Target I_t
Depth Net
Coarse Depth D_t^c
Reference & Target I_r & I_t
Flow Net
Flow $f_{r \leftarrow t}$
Pose Net
Pose $[R|t]_{rt}$
Static Background $\notin M_m^r$
New Reference $\hat{I}_r$
Dynamic Object $\in M_m^r$
Reference & Target I_r & I_t
Volume $\mathcal{V}_m$
Attention-Based Depth Net
Final Depth D_t

Figure 10: Illustration for the structure of Manydepth2.

3.4.3 Performance Table

Dataset	Abs Rel	Sq Rel	RMSE	RMSE log
KITTI	0.107	0.850	4.720	0.185

Table 4: ManyDepth2 performance.

3.5 Comparative Analysis

Model	Year	Abs Rel (KITTI)	δ < 1.25 (KITTI)	RMSE	Dataset	Type
Monodepth2	2019	0.115	87.4%	4.863	KITTI	Self-supervised
AdaBins	2021	0.103	89.9%	0.364	NYUv2	Supervised
BinsFormer	2022	0.099	91.2%	0.329	NYUv2	Supervised
ManyDepth2	2023	0.101	89.6%	0.328	KITTI	Self-supervised

Table 5: Comparative analysis of all four models.

3.6 Datasets Used

3.6.1 KITTI Dataset

- **Purpose**: Autonomous driving scenarios.
- **Content**: Stereo images with sparse LiDAR depth ground truth.
- **Usage**: Extensively used in Monodepth2 and ManyDepth2.
- **Resolution**: 1242x375.
- **Challenges**: Sparse depth, limited scene diversity.

3.6.2 NYU Depth v2

- **Purpose**: Indoor depth estimation.
- **Content**: RGB-D pairs captured using Microsoft Kinect in indoor settings.
- **Usage**: AdaBins and BinsFormer evaluated on this dataset.
- **Resolution**: 640x480.
- **Challenges**: Occlusion, indoor lighting, cluttered scenes.

CHAPTER 4: GAPS IN STUDY

Monocular Depth Estimation has shown significant promise, but several gaps and challenges still exist:

- **Limited Generalization:** Many models are trained and evaluated on structured datasets such as KITTI or NYU Depth V2, limiting their generalizability to unconstrained environments.
- **Dynamic Scenes:** Handling moving objects and non-rigid motion remains a major hurdle, especially in self-supervised setups.
- **Textureless and Reflective Surfaces:** Depth estimation struggles on surfaces that lack texture or have high reflectivity, due to poor visual cues.
- **Scale Ambiguity:** Monocular estimation inherently lacks absolute scale, requiring additional supervision or assumptions.
- **Computational Load:** Many accurate models are computationally intensive and not suited for real-time or embedded systems.
- **Data Dependency:** Supervised models rely heavily on ground truth depth data, which is costly to obtain and often sparse or noisy.
- **Temporal Inconsistency:** In video sequences, the lack of temporal consistency in predictions leads to flickering and unreliable depth maps.

CHAPTER 5: APPLICATIONS

- **Autonomous Driving:** Perception systems in self-driving cars use MDE for obstacle avoidance, navigation, and scene understanding.
- **Robotics:** Indoor and outdoor robots use MDE to perceive and interact with their environment without expensive sensors.
- **Augmented Reality (AR):** Realistic object placement and interaction in AR applications are enabled by depth awareness.
- **3D Scene Reconstruction:** MDE helps reconstruct dense 3D models from simple video or image sequences.
- **Medical Imaging:** In endoscopy and other visual diagnostics, depth information improves scene comprehension.
- **Virtual Reality and Gaming:** Realistic rendering and physics simulations are possible with depth information from 2D input.

CHAPTER 6: CONCLUSION

Monocular depth estimation has evolved rapidly due to advances in deep learning. From early CNN-based models like Monodepth to modern hybrid models using transformers and self-supervised learning, the field continues to address challenges around generalization, speed, and real-world deployment. While each model reviewed here brings unique strengths, limitations such as dynamic scene handling and scale ambiguity remain. Continuous innovation in architecture design, loss functions, and training paradigms is necessary to realize robust, real-time depth estimation from a single image.

CHAPTER 7: FUTURE WORK

- **Lightweight and Efficient Architectures:** Future models could integrate efficient transformer variants or lightweight CNN-transformer hybrids.
- **Improved Dynamic Scene Understanding:** Incorporate motion cues and object tracking. Example: ManyDepth2 combines optical flow with self-supervised loss.
- **Cross-Domain Generalization:** Use domain adaptation (e.g., CycleGAN, style transfer) to adapt models without retraining.
- **Self-Supervised Learning Enhancements:** Use photometric loss, temporal geometric constraints, and occlusion reasoning.
- **Temporal Consistency:** Employ LSTM units, 3D CNNs, or temporal attention to ensure smooth predictions across frames.
- **Uncertainty Estimation:** Use Bayesian networks or dropout for confidence-aware predictions.
- **Multi-modal Fusion:** Integrate RGB with event camera or IMU data for robust predictions.
- **Benchmarking:** Evaluate across synthetic and real-world datasets to ensure real-world deployment readiness.

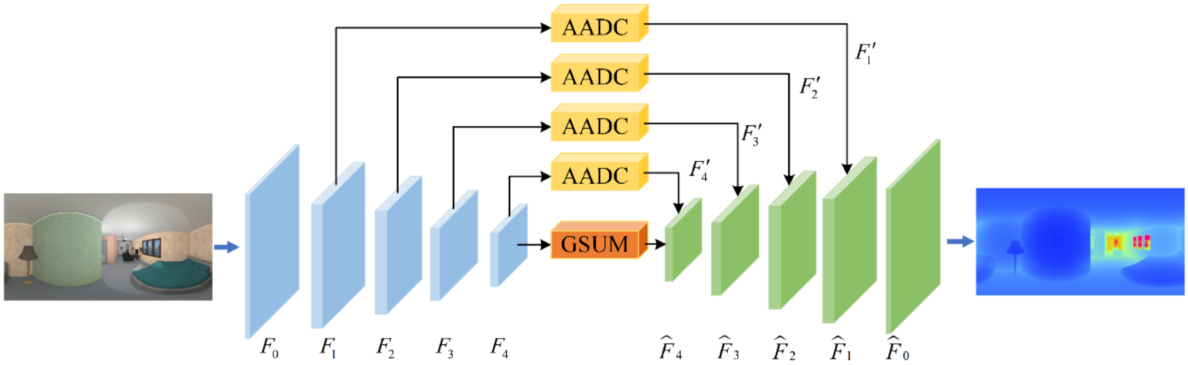

Figure 11: Lightweight depth network example.

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