CYSHIELD

Depth Estimation Evaluation Report

Abstract

We evaluated three monocular depth estimation models (DPT-Hybrid, MiDaS_small, ZoeDepth_N) and two stereo methods (RAFT-Stereo, SGBM) on a 50-image subset of the CARLA dataset. Quantitative evaluation covered accuracy metrics (AbsRel, RMSE, MAE, SILog, δ-thresholds, EPE) and efficiency (inference time). RAFT-Stereo achieved the best accuracy, with **AbsRel = 0.0874, RMSE = 9.55 m, δ<1.25 = 0.95** at ~15 ms per image, while SGBM performed poorly (AbsRel > 81, RMSE > 960 m). Among monocular methods, DPT-Hybrid was the strongest (**AbsRel = 0.289, RMSE = 16.86 m**) but still lagged behind stereo, while MiDaS_small offered speed advantages (0.16 ms) at the cost of higher errors. These results highlight that learning-based stereo remains superior for metric depth, whereas monocular models trade accuracy for efficiency and robustness in low-texture scenes.

Introduction

Depth estimation is a core task in 3D computer vision, essential for robotics, AR/VR, and autonomous driving. Two major approaches exist:

- **Monocular depth estimation**: Predict depth from a single RGB image using learned priors. Advantage: it works with one camera.

  Limitation: suffers from scale ambiguity.

- **Stereo depth estimation**: Predict depth from a calibrated stereo pair by computing disparity. Advantage: produces metric depth.

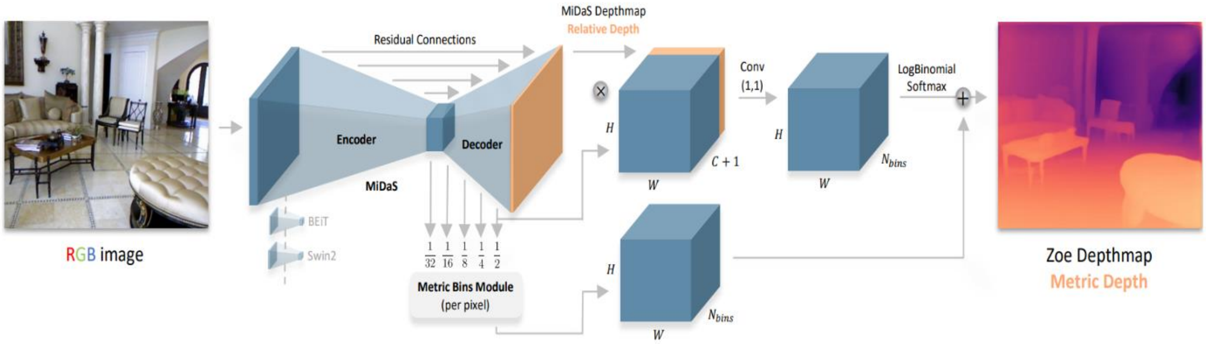
  Limitation: fails in textureless or reflective regions.

In this work, we systematically evaluate monocular and stereo methods on the CARLA dataset, comparing their accuracy, efficiency, and robustness.

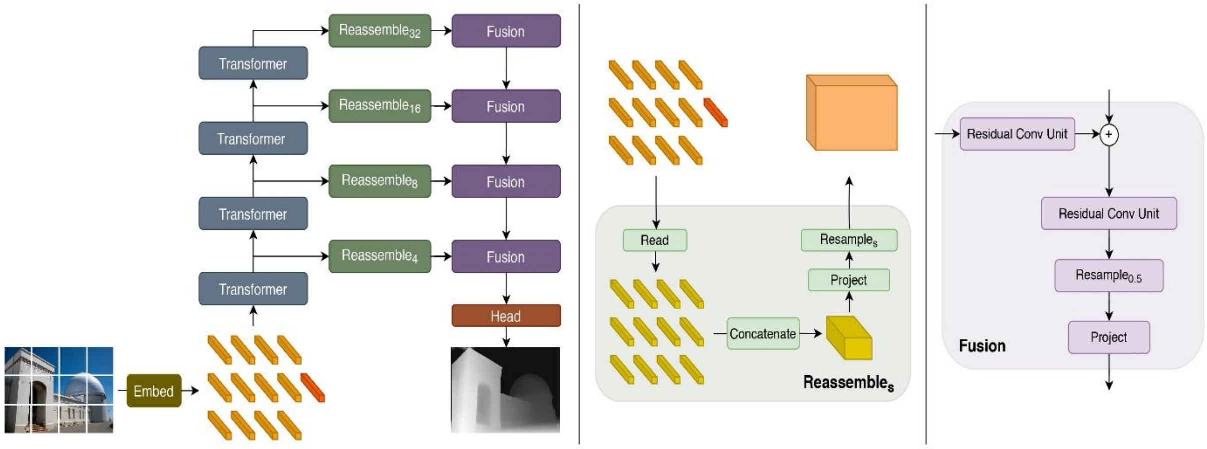
Methods

Models

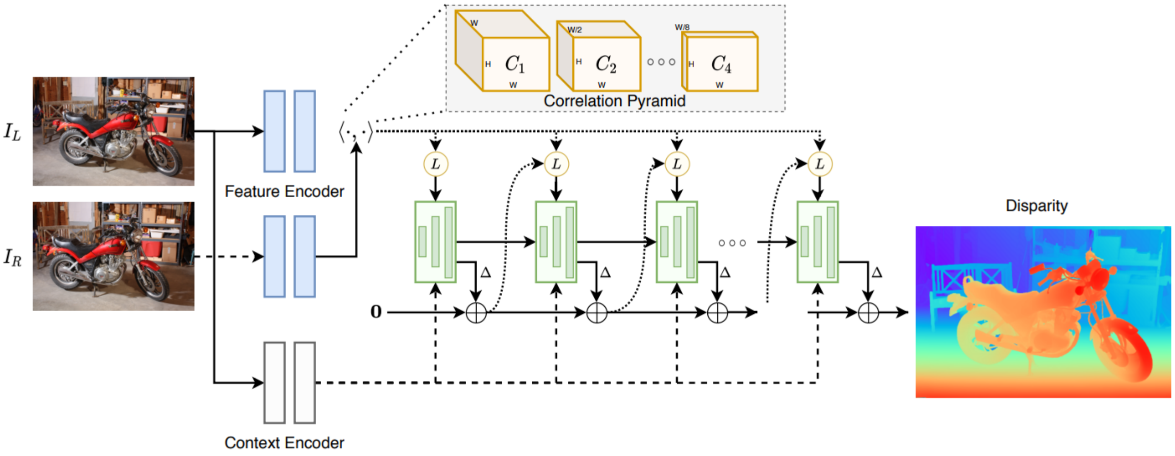
- Monocular:
  - *MiDaS*: multi-task, trained on mixed datasets
  - *Zoe*: ZoeDepth extends the MiDaS family by introducing scale-invariant depth supervision and improved normalization to handle domain shifts and predicts metric depth maps directly.



- *DPT_hybrid*: Vision Transformer backbone combined with CNN components to capture both global context and local spatial detail.



- Stereo:
  - *RAFT-Stereo* constructs a dense 4D correlation volume between image pairs and refines disparity estimates. Its architecture emphasizes all-pairs matching and recurrent refinement, enabling precise disparity estimation even in challenging regions.



  - *SGBM*: classical stereo algorithm that computes block-matching costs along multiple scanlines and then applies semi-global optimization to enforce smoothness.

Metrics

We report:

- **AbsRel** = mean absolute relative error
- **RMSE**, **MAE**, **RMSElog**
- **SILog** (scale-invariant log error)
- **$\delta<1.25$, $\delta<1.25^2$, $\delta<1.25^3$** (threshold accuracy)
- **EPE** (End-Point Error in disparity, for stereo only)
- **Inference time** (ms/image)

Evaluation Protocol

- **Dataset**: 50 CARLA images with ground-truth depth.
- **Invalid pixels:** After ground truth analysis, we found that there is discontinuity in pixels after 164 so it took as a threshold for valid mask. 0 and >164 are masked out.
- **Depth range:** 1–164 m.
- Monocular predictions aligned using **median scaling.**
- Stereo depth from disparity: depth = f * B / disparity.
  Where f = 658.5570007600752 (focal length in mm).
  B = 0.4 (Baseline in meters)
- **Hardware & Environment:** For stereo NVIDIA Tesla T4 GPU (16 GB memory) provided by Colab. And Intel Xeon CPU @ 2.20 GHz with 2 vCPUs for Monocular.

Results

Quantitative results

Model	AbsRel	RMSE	MAE	SILog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Infer-time
Midas	0.478	19.357	10.216	0.495	0.337	0.614	0.791	0.156
Zoe	1.154	20.743	12.462	0.682	0.218	0.335	0.407	13.546
DPT-Hybrid	0.288	16.855	8.440	0.347	0.488	0.788	0.907	3.421
RAFT-Stereo	0.087	9.552	2.555	0.333	0.945	0.957	0.964	14.838
SGBM	81.2182	967.887	360.756	2.070	0.817	0.835	0.842	0.749

Summary table above presents the averaged results across all test images for monocular and stereo models. **RAFT-Stereo clearly outperformed all other methods**, achieving the lowest errors (AbsRel = **0.087**, RMSE = **9.55 m**, MAE = **2.56 m**, SILog = **0.333**) and the highest accuracy thresholds (δ<1.25 = **0.945**). This confirms its ability to produce reliable metric depth while maintaining a reasonable inference time (~15 ms).

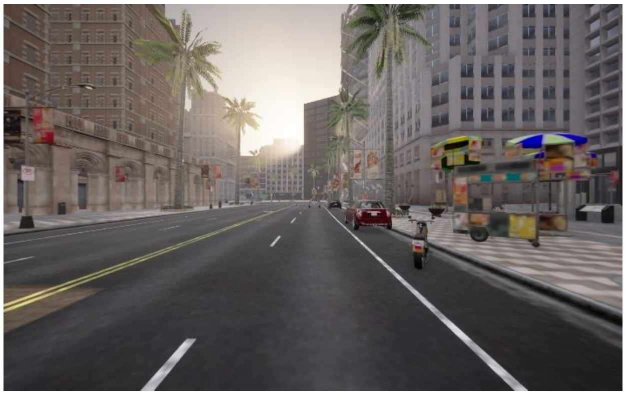
Among monocular approaches, **DPT-Hybrid was the strongest**, with AbsRel = **0.289** and δ<1.25 = **0.489**, substantially better than MiDaS_small and ZoeDepth_N. While MiDaS_small was extremely fast (0.16 ms), it traded accuracy (AbsRel = **0.478**), making it suitable only for lightweight applications. ZoeDepth_N struggled the most, with AbsRel exceeding **1.15** and very low δ accuracies, suggesting weak generalization in the CARLA domain.

The classical **SGBM baseline performed poorly**, with errors an order of magnitude larger than all other methods (AbsRel > **81**, RMSE ≈ **968 m**), showing its limitations in synthetic driving environments with large depth ranges.

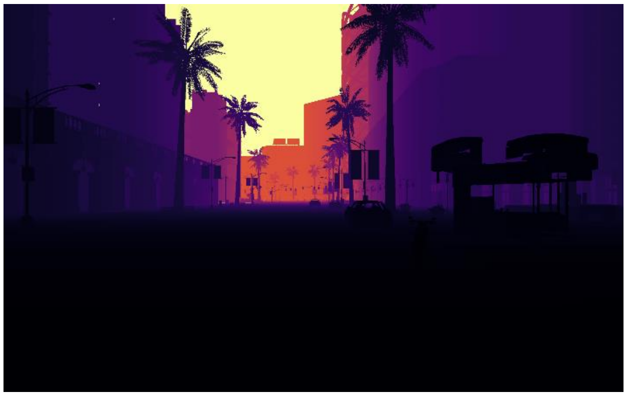
In summary, the table highlights a **clear trade-off between stereo and monocular approaches**: stereo (RAFT-Stereo) delivers the most accurate metric depth, while monocular methods provide faster inference with reduced accuracy.

Qualitative results

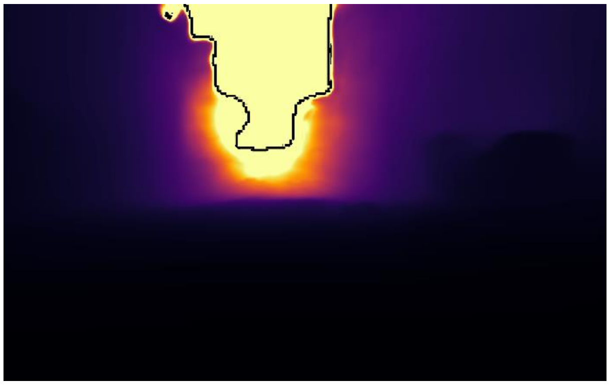
Input Image



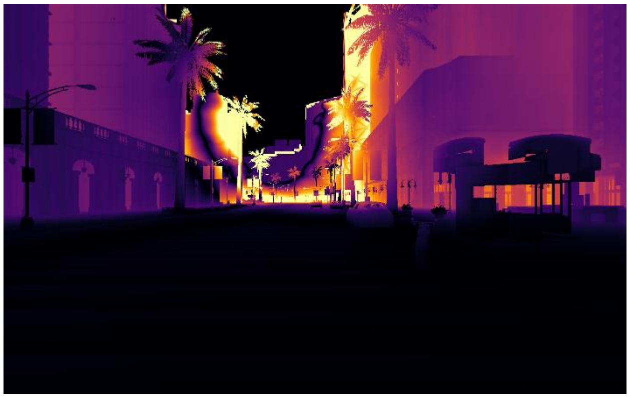
Ground truth



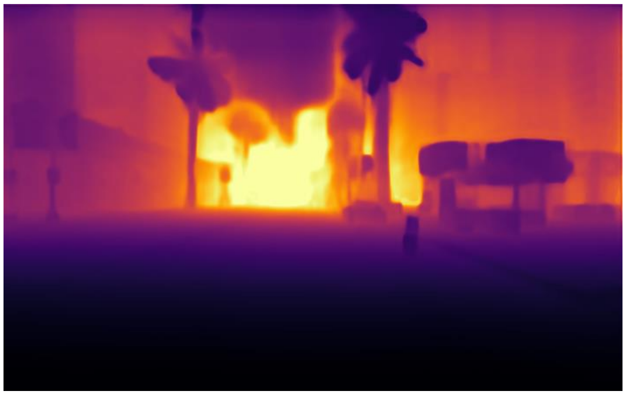
Midas Depth map



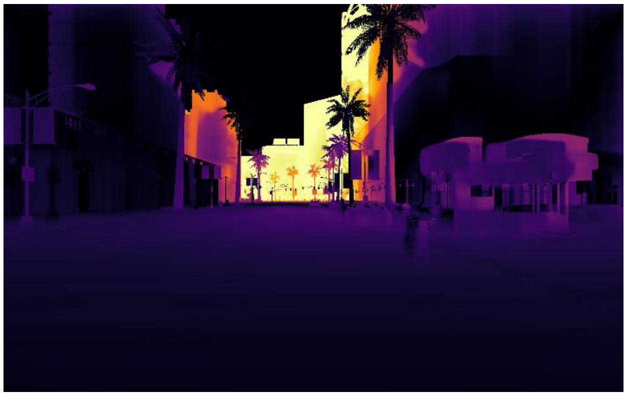
Midas Error map



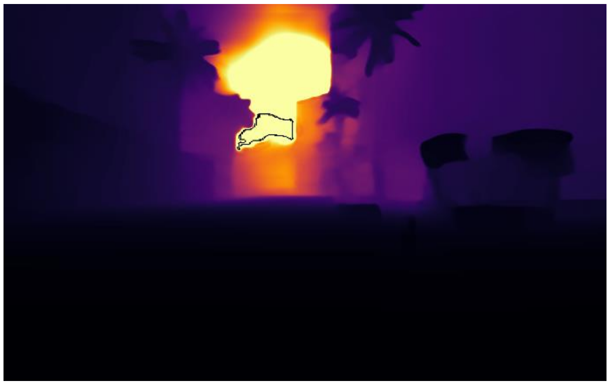
Zoe Depth map



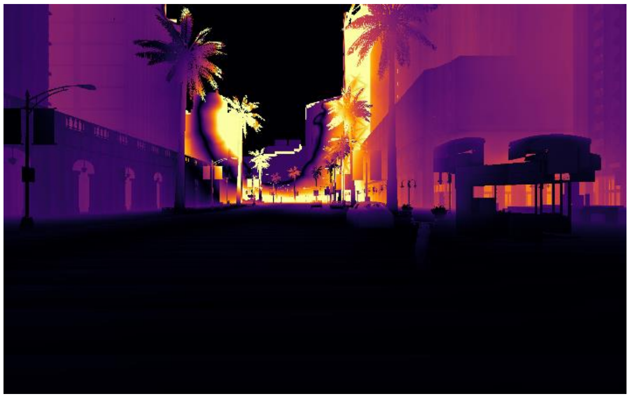
Zoe error map



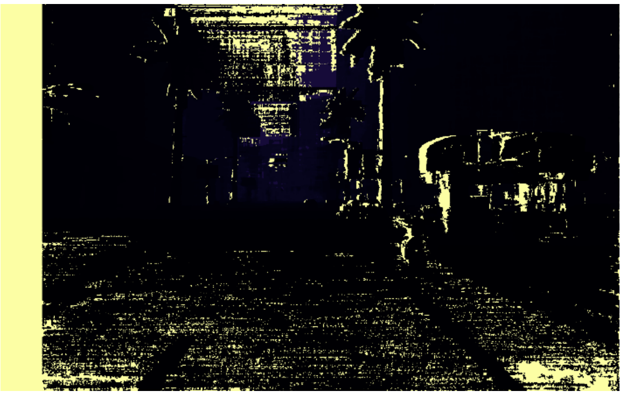
DPT Hybrid depth map



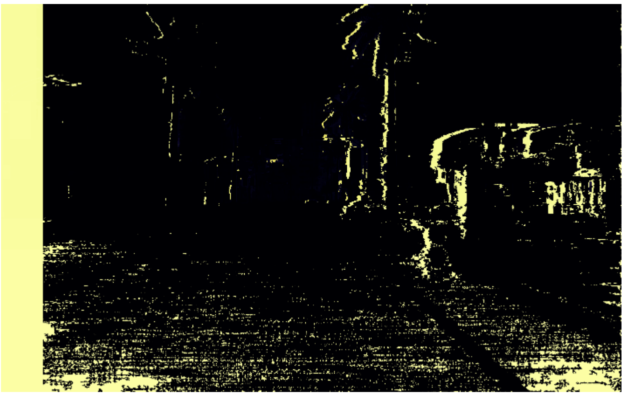
DPT Hybrid error map



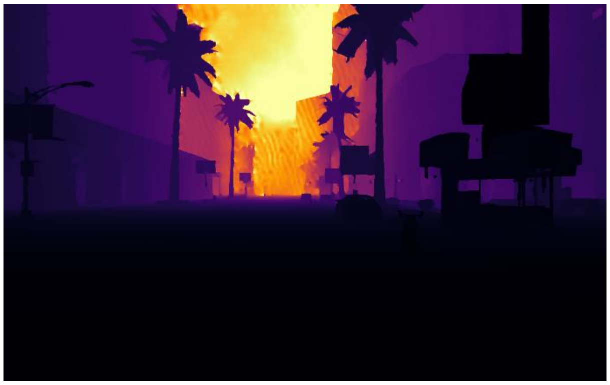
SGBM Depth map



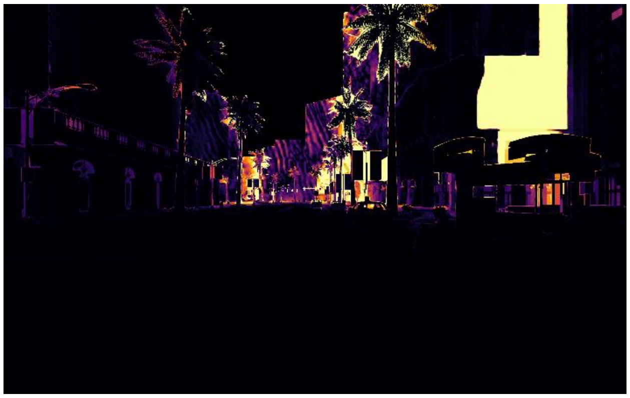
SGBM error map



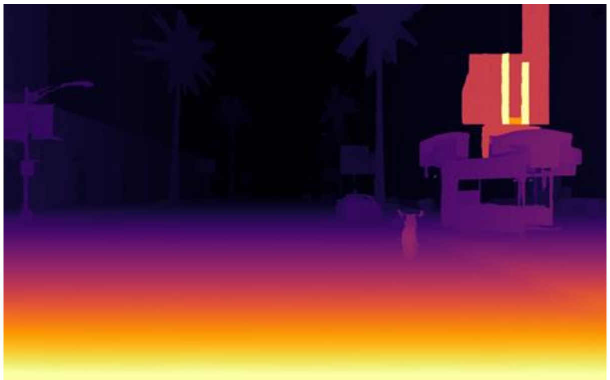
RAFT-Stereo depth map



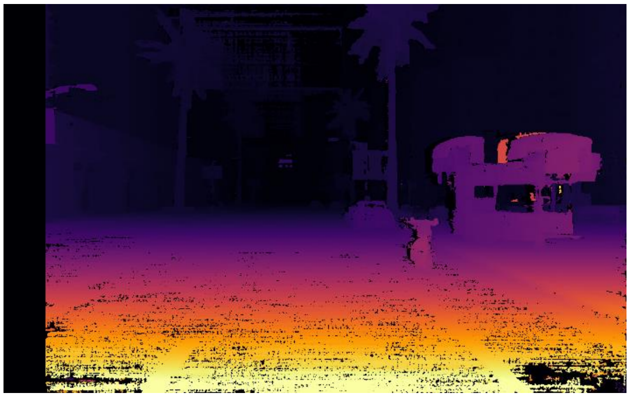
RAFT-Stereo error map



RAFT-Stereo disparity



SGBM disparity



Discussion

- RAFT-Stereo
    - **Strengths:** State-of-the-art accuracy, recovering fine geometric details and producing reliable metric depth. Supports "anytime prediction" with its iterative refinement.
    - **Weaknesses:** Computationally heavier than monocular models; struggles with occlusions, ill-posed regions, and large texture-less or reflective surfaces where stereo matching becomes ambiguous.

  - **Failure Cases:** Disparity holes in texture-less or reflective regions, as stereo matching relies on visual correspondence which is ambiguous in these areas.
  - **Potential Improvement:** Incorporating monocular priors or post-processing to fill disparity holes. Using LiDAR guidance and pre-filling sparse data can also improve performance in challenging scenes.
- SGBM
  - **Strengths:** Computationally efficient on CPU (~0.75 ms), easy to implement as a classic baseline.

  - **Weaknesses:** Extremely poor accuracy, especially in synthetic data with large depth ranges. Relies on hand-crafted features which limit its performance.

  - **Failure Cases:** Fails completely in distant regions and complex structures, producing unreliable disparity maps. Suffers from blurred object boundaries and fine details.

  - **Potential Improvement:** Parameter tuning, cost filtering, or hybridizing with learning-based post-processing. Modern approaches with deep learning can provide much smoother and more consistent depth maps.

- DPT-Hybrid
  - **Strengths:** The strongest monocular method, balancing a transformer-based global context with CNN local detail. Achieves state-of-the-art results on specific datasets like KITTI.
  - **Weaknesses:** Significantly less accurate than stereo methods; produces scale drift and over-smoothing in fine structures due to the inherent ill-posed nature of monocular depth estimation.

    - **Failure Cases:** Struggles with reflective surfaces and very distant objects where depth cues are insufficient.
    - **Potential Improvement:** Fine-tuning on domain-specific data (e.g., CARLA or other synthetic driving datasets) to improve adaptation. Leveraging a three-stage pipeline to address scale ambiguity.
- Midas
    - **Strengths:** Extremely fast (0.16 ms) and lightweight, suitable for real-time or resource-constrained environments.
    - **Weaknesses:** Accuracy is significantly lower, reflecting the trade-off for speed. Estimated depth maps are often low-resolution and lack fine details.
    - **Failure Cases:** Produces blurry, low-detail depth maps, particularly failing to capture thin structures.
    - **Potential Improvement:** Knowledge distillation from larger MiDaS variants to improve accuracy while retaining speed. The original paper demonstrated that combining multiple datasets for training can lead to superior zero-shot performance.
- ZoeDepth
    - **Strengths:** Designed for generalization across datasets. When the evaluation depth range was increased from 164 m to 200 m, its performance improved, suggesting its scale handling benefits from broader ranges.
    - **Weaknesses:** In the standard 164m setup, it severely underperformed, failing to produce reliable depth at closer ranges. This model is sensitive to the range and scale of the data.
    - **Failure Cases:** Failed to produce reliable depth at closer ranges, often mis-scaling objects in the scene.
    - **Potential Improvement:** Domain-specific fine-tuning or adaptive scaling strategies to stabilize predictions. The original paper proposes a novel bin adjustment design and a framework that can jointly train on multiple datasets to improve performance