Enhancing Monocular Metric Depth Estimation through Adaptive Scaling

Tae Yang [1]

[1]taeyang@stanford.edu

Introduction

Monocular Depth Estimation (MDE) is widely used in robotics, AR, and autonomous systems, yet predicting **accurate metric depth** remains a challenge due to **scale ambiguity**, where absolute depth information is lost in projection.

Limitations of Existing Methods:

- **Relative depth only** – Most models predict ordinal depth but fail to recover true metric scale.
- **Sensor dependency** – Methods using LiDAR/stereo improve accuracy but increase cost and complexity.
- **Fixed depth constraints** – Some models assume a predefined maximum depth, limiting adaptability.

Proposed Approach: Adaptive Depth Scaling

- **Dynamically corrects scale errors** by predicting an image-specific depth scaling factor.
- **Two-step solution:**
  - Generate a dataset of **optimal scaling factors** by minimizing scale-invariant errors.
  - Train a **lightweight CNN** to infer depth scaling factors from input images.

Data Generation: Learning Optimal Scaling Factors

Addressing Scale Ambiguity:

- **SOTA MDE models** accurately predict **relative depth** but exhibit **systematic scale errors**.
- Analyzing their **predictions vs. ground truth** reveals that **correcting scale improves metric depth estimation**.

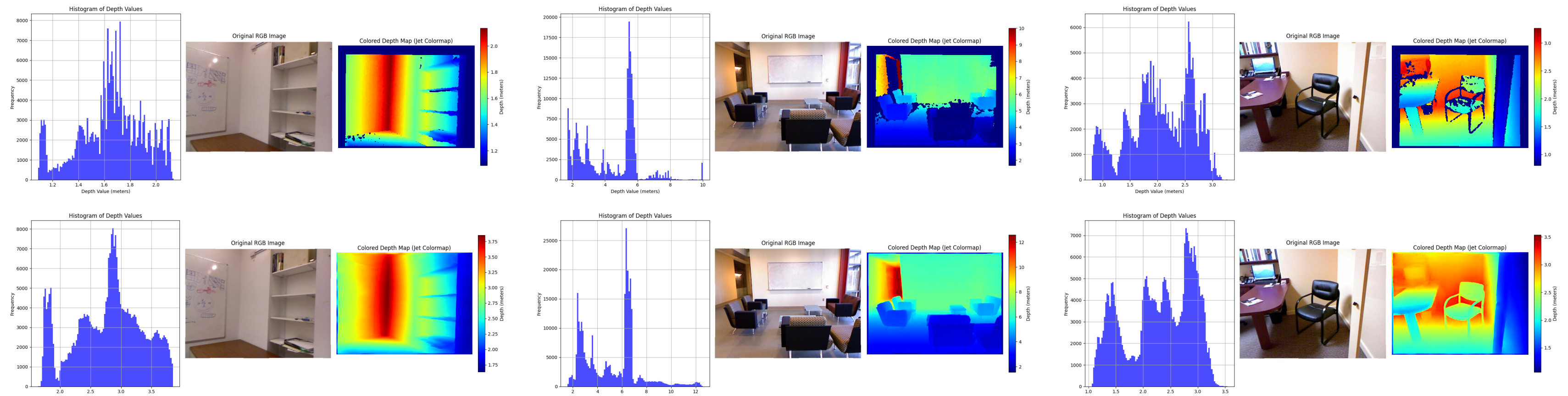


Figure 1. Comparison of ground truth depth (top row) and Depth Anything V2 predictions (bottom row). Observing scale errors in these predictions motivated our scaling optimization approach.

- We generate **image-specific depth scaling factors** to refine these predictions.

Optimal Scaling Factor Computation:

- Given a predicted depth map, we determine the optimal scaling factor by minimizing **scale-invariant errors**.
- We use the **Wasserstein distance** to align predicted and ground-truth depth distributions.

Optimization Objective:

$$s^* = \underset{s \in [0.1, 2.0]}{\arg\min}\ W(s \cdot D_{\text{pred}}, D_{\text{gt}})$$

where W is the Wasserstein distance, D_{pred} is the predicted depth, and D_{gt} is the ground truth depth.

Building a Training Dataset:

- **392 images** from NYU Depth V2 used to extract optimal scaling factors.
- **Per-image scaling factors** are log-transformed for numerical stability.
- The dataset is used to train a CNN for automatic depth scale correction.

Training: Learning Depth Scaling Factors

Neural Network Architecture:

- **Lightweight CNN** predicting log-transformed depth scaling factor.
- Two convolutional layers (64, 128 channels) with ReLU and MaxPooling.
- Flattened output passed through a 512-unit fully connected layer.
- Final linear layer outputs a **single scale factor** per image.

Training Process:

- Dataset: 392 images from NYU Depth V2, each with RGB input and scale factor.
- Loss Function: Mean Squared Error (MSE).
- Optimizer: Adam ($lr = 10^{-3}$), step decay every 10 epochs.
- Training for 30 epochs, validating every 5 epochs.
- Batch size = 30.

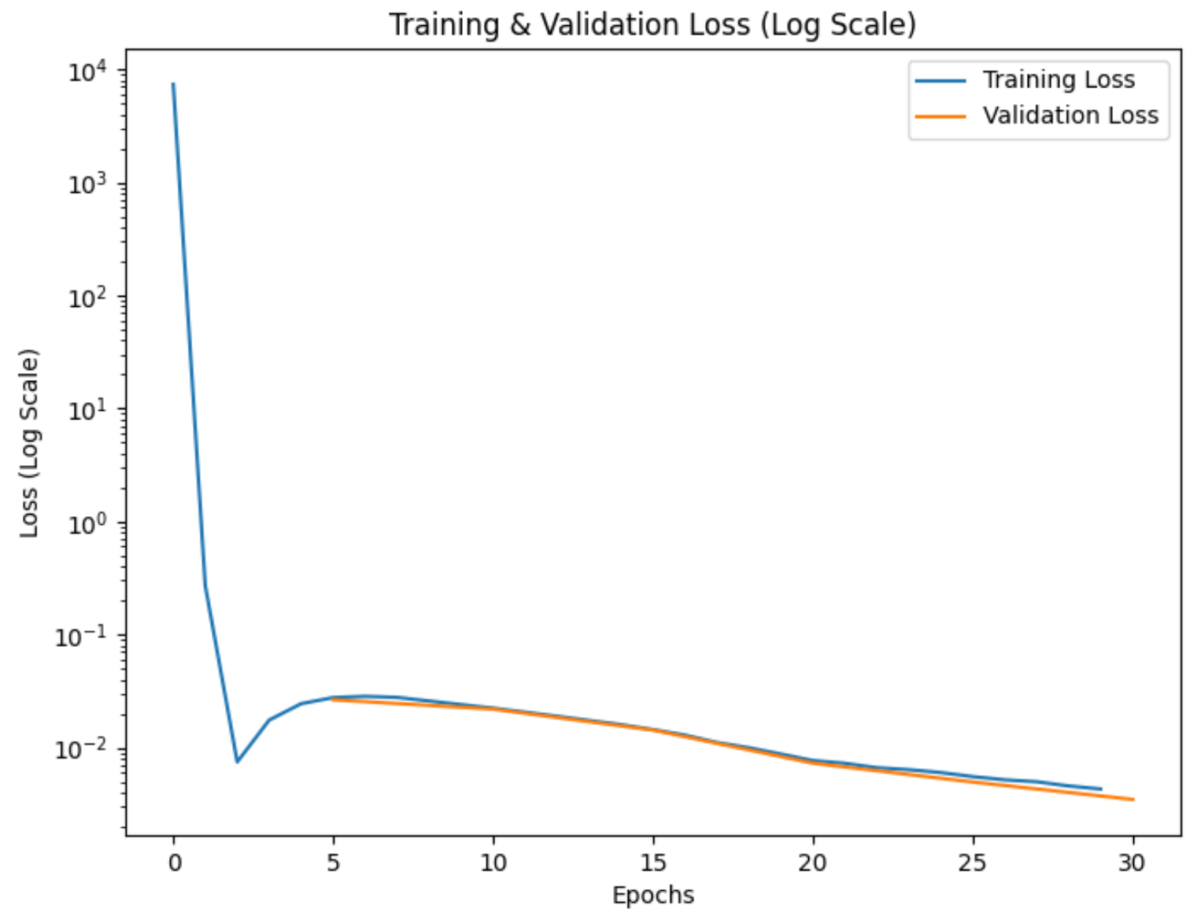


Figure 2. Training and Validation Loss (Log Scale).

Quantitative Evaluation: Depth Estimation Accuracy

Performance Metrics:

- **AbsRel (Absolute Relative Error):** Measures depth estimation accuracy.
- **RMSE (Root Mean Squared Error):** Penalizes large depth errors.
- **δ_1 Accuracy:** Measures percentage of correctly estimated depths.
- **Scaling Factor Estimation:** Evaluates accuracy of predicted depth scale factors.

Table 1. Scaling Factor Estimation Performance.

Dataset	MAE ↓	RMSE ↓
Training Set	0.0439	0.0532
Test Set	0.1172	0.1269

Key Findings:

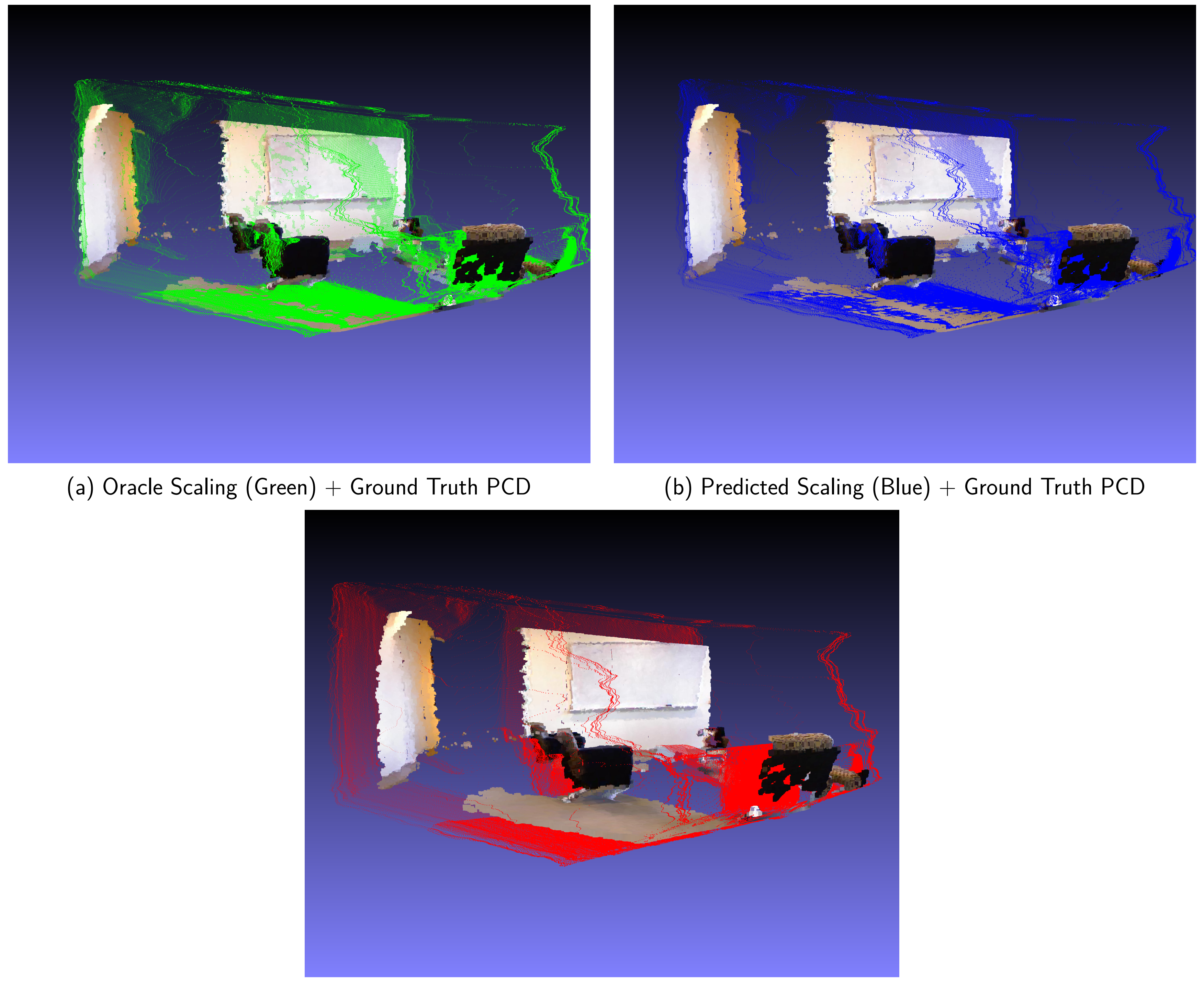
- Our **adaptive scaling** significantly improves depth estimation accuracy.
- **Scaling factor prediction is reliable**, with low MAE and RMSE values.
- **Our method generalizes well** across test scenes, reducing errors without manual tuning.
- **Achieves performance close to oracle scaling**, proving the effectiveness of learned depth correction.

Table 2. Depth Estimation Performance on Training and Test Sets.

Method	AbsRel ↓	RMSE ↓	δ_1 (%) ↑
Training Set			
DA2 w/o scaling	0.2219	0.9480	66.82
DA2 w/ Oracle Scale	**0.0554**	**0.3083**	93.87
DA2 w/ Adaptive Scale (ours)	0.0792	0.3632	**96.07**
Test Set			
DA2 w/o scaling	0.3380	0.6166	32.06
DA2 w/ Oracle Scale	**0.1084**	**0.2307**	**87.94**
DA2 w/ Adaptive Scale (ours)	0.1914	0.3636	72.31

Qualitative Evaluation: 3D Point Cloud Comparisons

3D Point Cloud Reconstruction: We visualize 3D point clouds generated from estimated depth maps to evaluate the effect of our adaptive scaling approach on monocular metric depth estimation. Below, we compare ground truth depth, oracle-scaled predictions, our adaptive scaling predictions, and unscaled DA2 predictions.



(a) Oracle Scaling (Green) + Ground Truth PCD

(b) Predicted Scaling (Blue) + Ground Truth PCD

(c) Unscaled DA2 (Red) + Ground Truth PCD

Figure 3. 3D Point Cloud Comparison: (a) Oracle-scaled DA2 prediction (green) aligns well with the ground truth, (b) our adaptive scaling (blue) significantly improves scale alignment, and (c) unscaled DA2 predictions (red) exhibit severe scale misalignment and distortions.

Conclusion and Future Work

Conclusion:

- Proposed a **learned adaptive scaling** approach to address **scale ambiguity** in monocular depth estimation.
- Experiments show **significant reduction in scale drift** without the help of external sensors, achieving near-oracle performance.
- Both **quantitative and qualitative analyses** confirm that adaptive scaling enhances depth accuracy and geometric consistency.

Future Work:

- Develop a mechanism to **detect erroneous depth predictions** before inference to improve reliability.
- Extend the approach to **pixel-wise adaptive scaling** for handling depth discontinuities and complex scene variations.