Enhancing Monocular Metric Depth Estimation through Adaptive Scaling

Tae Yang[1]

Abstract— **Monocular metric depth estimation remains a challenging problem due to scale ambiguity, which limits the ability of models to produce accurate metric depth predictions across diverse scenes, even though they excel at predicting relative depth. While recent models have been fine-tuned for improved metric depth estimation, they still experience systematic scale inconsistencies across different datasets and scenes. In this work, we introduce an adaptive depth scaling framework that predicts an image-specific depth scaling factor, improving metric depth estimation accuracy without requiring scene-dependent manual tuning. We first generate a dataset of optimal depth scaling factors by minimizing scale-invariant errors between predicted and ground-truth depth maps. A lightweight convolutional neural network (CNN) is then trained to predict these scaling factors directly from input images, allowing for dynamic correction of monocular depth predictions. We evaluate our method on the NYU Depth V2 dataset, demonstrating that our approach outperforms baselines and enhances depth estimation accuracy across various evaluation metrics. This work highlights the importance of image-conditioned scale adaptation and contributes toward more robust depth estimation for applications such as robotics, augmented reality (AR), and scene understanding.**

I. Introduction

Monocular depth estimation (MDE) is a fundamental task in computer vision with applications in autonomous driving, robotics, and augmented reality (AR) [1], [2]. Despite significant advancements in deep learning-based approaches, accurately predicting metric depth from a single image remains challenging due to *scale ambiguity*, which arises from the loss of absolute scale information during perspective projection [3], [4]. This ambiguity hinders real-world applications that require precise spatial measurements, such as robot navigation and scene reconstruction [5].

Traditional MDE models often focus on relative depth estimation, learning ordinal relationships between objects in a scene without predicting absolute distances [6]. While such approaches achieve high accuracy in ranking depth relationships, they fail to generalize across diverse environments where depth scale varies significantly. Some methods attempt to resolve this by incorporating auxiliary sensors such as LiDAR [7] or stereo images [8], but these solutions increase hardware complexity and cost. Others leverage self-supervised learning with geometric constraints [9] or predefine a fixed maximum depth assumption to constrain predictions [10], yet these approaches fail to generalize across different datasets and scenes.

[1]T. Yang is with the Department of Computer Science, Stanford University, CA, USA taeyang@stanford.edu

Recent studies have attempted to bridge the gap between relative and metric depth by introducing learned scaling factors [10], scene-dependent depth priors [11], and velocity guidance for dynamic scale estimation [12]. However, many of these approaches still struggle with *scalability and adaptability* when deployed across varying environments.

To address these limitations, we propose a learned adaptive depth scaling framework that dynamically predicts an **image-specific depth scaling factor**, improving the accuracy of monocular metric depth estimation without requiring additional sensors or manual tuning. Our approach consists of two key steps:

- Generating a dataset of **optimal depth scaling factors** by minimizing **scale-invariant errors** between predicted and ground-truth depth maps from the NYU Depth V2 dataset [13].
- Training a **lightweight convolutional neural network (CNN)** to infer depth scaling factors directly from input images, allowing for **dynamic correction** of MDE outputs.

We evaluate our method against state-of-the-art (SOTA) monocular depth estimation models, demonstrating significant improvements across multiple evaluation metrics, including **absolute relative error (AbsRel), root mean squared error (RMSE), and δ_1 accuracy**. By learning an **image-conditioned scaling factor**, our approach enhances the reliability of **monocular metric depth estimation (MDE)** in scale-sensitive applications such as robotics, augmented reality (AR), and scene understanding.

II. Related Work

Existing MDE methods can be primarily categorized based on their strategies to address scale ambiguity:

Supervised Depth Regression. Early seminal works by Eigen et al. [3] pioneered direct metric depth prediction from RGB images using multi-scale convolutional neural networks (CNNs). Subsequent approaches, such as those by Fu et al. [14], leveraged ordinal regression for improved depth accuracy but still faced scale ambiguity, resulting in inconsistencies across datasets and scenes.

Self-Supervised and Geometric Constraints. To circumvent the dependency on extensive labeled datasets, self-supervised methods have exploited geometric constraints, for instance, photometric consistency and multi-view reconstruction losses [5], [11]. Although self-supervision avoids explicit reliance on ground truth depth data, it struggles with scale generalization, typically necessitating manual scale calibration per dataset or environment [11].

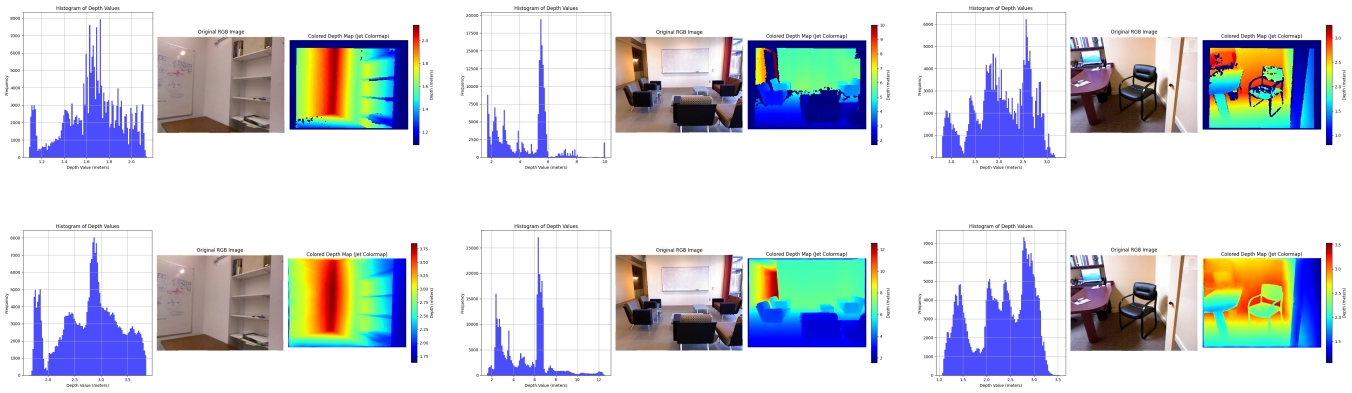


Fig. 1: Depth distributions, RGB images, and corresponding depth maps from Ground Truth and Depth Anything V2 predictions.

Relative Depth Approaches with Post-hoc Scaling. Models like MiDaS [6], ZoeDepth [2], and Depth Anything V2 [10] initially predict relative depth, achieving high accuracy in determining ordinal relationships between pixels without absolute distance. To obtain metric depth, these approaches require dataset-specific scaling, typically achieved through manual or heuristic tuning. This reliance on fixed or manually set scale parameters significantly limits their adaptability to unseen datasets or environments.

Depth Anything V2 [10] currently represents the state-of-the-art in monocular depth estimation, achieving exceptional performance in terms of both accuracy and inference speed. Despite its strong performance in estimating relative depth, its fine-tuned metric predictions often exhibit scale deviations from ground truth.

Our work specifically addresses this limitation by learning an adaptive, image-conditioned scaling factor to further enhance depth estimates into reliable metric predictions. This approach directly targets the primary drawback identified in existing literature—the difficulty in generalizing scale across varying scenes and datasets—without requiring additional sensors or complex pre-processing steps.

III. Method

We first examine Depth Anything V2 [10], a state-of-the-art monocular depth estimation model, on the NYU Depth V2 dataset to assess its capabilities and limitations in metric depth estimation. As illustrated in Fig.1, the pretrained model generates depth predictions whose distributions closely resemble the ground truth distributions, suggesting accurate estimation of relative depths. However, despite this similarity in relative depth, there remains a consistent linear scale discrepancy between predicted and ground truth depths, indicating scale ambiguity inherent to monocular depth estimation. Thus, we hypothesized that identifying the appropriate image-specific scaling factor could substantially refine the metric depth estimates.

To accurately determine the optimal scaling factors, we devised an optimization-based approach grounded in aligning depth distributions. Given an input RGB image and its corresponding predicted depth map from Depth Anything V2, our goal is to find a scaling factor that minimizes the discrepancy between the predicted and ground truth depth distributions. To quantify this distributional difference, we employ the Wasserstein distance, also known as Earth Mover's Distance (EMD). Intuitively, Wasserstein distance measures the minimum amount of effort required to transform one probability distribution into another, where effort is quantified as the amount of "earth" moved multiplied by the distance moved [15], [16].

Formally, we determine the optimal scaling factor through a bounded optimization problem:

$$s^* = \underset{s\in[0.1,2.0]}{\arg\min}\, W(s \cdot D_{pred}, D_{gt}), \tag{1}$$

where W is the Wasserstein distance between the scaled predicted depth values ($s \cdot D_{pred}$) and the ground truth depth values (D_{gt}). This optimization problem is solved numerically using a bounded optimization solver, initialized at $s = 1.0$ and constrained within a predefined range for numerical stability and generalization. By solving this optimization problem individually per image, we generate accurate per-image scaling factors, thus enabling effective metric depth refinement.

Ultimately, our goal is to automate the estimation of optimal scaling factors directly from input RGB images. To systematically achieve this, we generate a supervised dataset by applying our optimization-based approach (Algorithm 1) to compute optimal scaling factors for a subset of 392 RGB-depth pairs from the NYU Depth V2 dataset [13]. Each scaling factor minimizes the Wasserstein distance between the scaled predicted depths and ground truth depths. The resulting dataset, comprising RGB images paired with their log-transformed optimal scaling factors, serves as training data for our neural network.

Using this generated dataset, we train a lightweight convolutional neural network (CNN) to predict the optimal scaling factor directly from an input RGB image (Algorithm 2). Notably, we found empirically that predicting the logarithm

Algorithm 1 Generating Dataset of Optimal Scaling Factors

Input: RGB images $\{X_i\}_{i=1}^N$, ground truth depths $\{D_{gt,i}\}_{i=1}^N$, predicted depths $\{D_{pred,i}\}_{i=1}^N$
for each RGB-depth pair $(X_i, D_{gt,i}, D_{pred,i})$ **do**
    Extract valid depth values: $D_{gt,i}^{valid}$, $D_{pred,i}^{valid}$
    Solve for optimal scaling factor:

$$s_i^* = \arg\min_{s\in[0.1,2.0]} W(s \cdot D_{pred,i}^{valid}, D_{gt,i}^{valid})$$

    Store the pair $(X_i, \log(s_i^*))$ in the dataset
end for
Dynamic Filtering (optional): Remove scaling factors outside 5th-95th percentile
Output: Dataset $\mathscr{D} = \{(X_i, \log(s_i^*))\}_{i=1}^{N'}$

Algorithm 2 Training CNN for Predicting Optimal Scaling Factors

Input: Training dataset $\mathscr{D} = \{(X_i, \log(s_i^*))\}_{i=1}^N$, learning rate η, epochs T
Initialize CNN model parameters θ, optimizer, learning rate scheduler
for epoch $= 1$ to T **do**
    for each batch $\mathscr{B} \subseteq \mathscr{D}$ **do**
        Predict log-scale factors: $\hat{y} = CNN(X_{\mathscr{B}}; \theta)$
        Compute MSE Loss: $\mathscr{L} = \frac{1}{|\mathscr{B}|}\sum_{j\in\mathscr{B}}(\hat{y}_j - \log(s_j^*))^2$
        Backpropagate gradients and update parameters θ
    end for
    Perform validation every 5 epochs, record validation loss
    Update learning rate scheduler
    Save model checkpoint periodically
end for
Output: CNN parameters θ^*

of the scaling factor stabilizes training by reducing the dynamic range and mitigating potential numerical instability caused by large variance in scale factors.

Specifically, the CNN architecture consists of sequential convolutional layers followed by fully connected layers, designed explicitly to capture global and local spatial structures. We hypothesize that the CNN implicitly leverages visual features and spatial contexts—such as edges, contours, textures, and semantic cues—in order to estimate the appropriate scaling factor without explicit geometric priors.

The CNN is trained to minimize the Mean Squared Error (MSE) loss between the predicted and ground-truth log-scale factors:

$$\mathscr{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{2}$$

where y_i denotes the ground-truth log-scale factor for the i-th image, and $\hat{y}_i$ is the predicted log-scale factor. Training uses the Adam optimizer with an initial learning rate of 10^{-3}, gradient clipping for stability, and a step-based learning rate scheduler that reduces the learning rate by half every 10 epochs.

Validation is performed periodically every five epochs to monitor the model's generalization capability and to mitigate potential overfitting over a held-out validation subset comprising 20% of the total dataset. The average validation loss is calculated using the same mean squared error (MSE) loss as during training.

Upon inference, the predicted log-scale is exponentiated to recover the actual scaling factor used for rescaling metric depth predictions from the pretrained Depth Anything V2 model.

IV. Experimental Results

We evaluate our learned adaptive scaling method by comparing its performance against baseline predictions from Depth Anything V2 (DA2) without a scaling factor, as well as predictions refined using the ground-truth optimal scaling factors computed via our optimization approach. Our analysis includes both quantitative and qualitative results, assessing the accuracy of scaling factor predictions and their impact on metric depth estimation performance. Specifically, we present: (1) training and validation loss curves to demonstrate model convergence; (2) evaluation of scaling factor predictions on training and test datasets; (3) comparative analyses of depth estimation accuracy (AbsRel, RMSE, δ_1) across methods; and (4) qualitative point-cloud visualizations illustrating improvements in depth estimation accuracy achieved through our adaptive scaling approach.

A. Training and Validation Loss

We first analyze our training and validation process by examining the training and validation loss curves (Fig. 2) where the losses are on a logarithmic scale for clarity. We observe a rapid initial decrease in both training and validation losses, indicating effective model learning during early training stages. Subsequently, the losses stabilize and gradually converge, suggesting the model reaches a stable optimum. Notably, the validation loss closely tracks the training loss, occasionally being slightly lower due to stochastic variations in data distribution during training, which suggests strong generalization capability without significant overfitting.

B. Quantitative Evaluation

To quantitatively assess the scaling factor estimation, we evaluate the trained model's performance in predicting scaling factors. Table I presents the results on both the training and unseen test datasets. The mean absolute error (MAE) and root mean squared error (RMSE) are reported as primary evaluation metrics to measure prediction accuracy.

TABLE I: Scaling Factor Estimation Performance

Dataset	MAE	RMSE
Training Set	0.0439	0.0532
Test Set	0.1172	0.1269

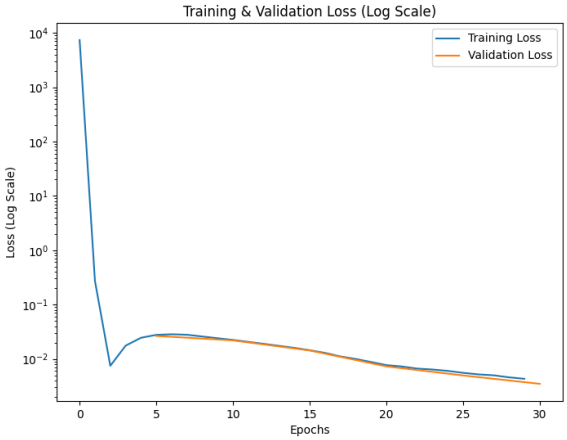


Fig. 2: Training and Validation Loss Curves

These results confirm that our lightweight CNN effectively learns to predict scaling factors, achieving low error rates on the training set while maintaining reasonable generalization to unseen test data. Although the test set errors are higher, this is expected due to variations in scene content and distribution shifts. The relatively low RMSE values suggest that the model captures the underlying relationship between image features and scale factors, supporting its potential for real-world applications in adaptive depth estimation.

Next, we evaluate our method's effectiveness in enhancing monocular metric depth estimation. Table II presents depth estimation performance in terms of absolute relative error (AbsRel), root mean squared error (RMSE), and accuracy threshold metric (δ_1), across three approaches: (1) DA2 without scaling, (2) DA2 with ground-truth optimal scaling (oracle), and (3) DA2 with learned adaptive scaling (our method). We report results separately for both the training and test sets.

To formally define these metrics:

- **Absolute Relative Error (AbsRel):** Measures the average relative error between predicted and ground truth depths, capturing the overall accuracy of depth estimation.

$$\text{AbsRel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i^{pred} - D_i^{gt}|}{D_i^{gt}} \tag{3}$$

- **Root Mean Squared Error (RMSE):** Evaluates the magnitude of depth estimation errors, penalizing larger discrepancies more heavily.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(D_i^{pred} - D_i^{gt})^2} \tag{4}$$

- **Accuracy Threshold Metric (δ):** Measures the percentage of predicted depths within a threshold ratio of the ground truth depths. We primarily report δ_1, defined as:

$$\delta_1 = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\left(\max\left(\frac{D_i^{pred}}{D_i^{gt}}, \frac{D_i^{gt}}{D_i^{pred}}\right) < 1.25\right) \tag{5}$$

where $\mathbb{1}(\cdot)$ is an indicator function that counts the fraction of pixels satisfying the threshold.

A lower AbsRel and RMSE indicate better depth estimation accuracy, while a higher δ_1 suggests improved consistency between predicted and ground truth depths.

TABLE II: Depth Estimation Performance on Training and Test Set

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

These results demonstrate that our learned adaptive scaling method substantially improves depth estimation accuracy over the baseline without scaling. On the training set, our approach achieves an AbsRel of 0.0792, which, while slightly worse than the oracle scale, still represents a large improvement over the baseline (0.2219). Similarly, on the test set, our method generalizes reasonably well, lowering AbsRel from 0.3380 (w/o scaling) to 0.1914. While the performance gap between the oracle and adaptive scaling increases in the test set, the significant improvement over the baseline model validates our hypothesis that learning an image-conditioned scaling factor enhances depth prediction accuracy across diverse scenes.

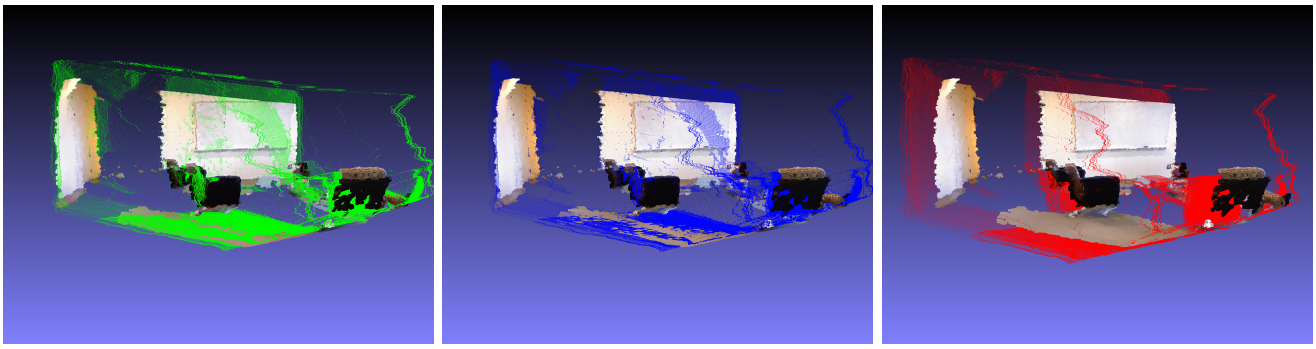
C. Qualitative Evaluation

To better illustrate the impact of our learned adaptive scaling, we visualize 3D point cloud reconstructions generated from different depth estimation strategies in Fig. 3. Each visualization includes the same RGB-colored point cloud derived from the ground truth depth map for reference, overlaid with an additional point cloud from one of the depth estimation methods:

- **Left**: The point cloud from DA2 predictions rescaled using oracle scaling factors, colored in green.
- **Middle**: The point cloud from DA2 predictions rescaled using our learned adaptive scaling, colored in blue.
- **Right**: The point cloud from DA2 baseline predictions without any scaling adjustment, colored in red.

The qualitative comparison clearly demonstrates the effectiveness of our learned adaptive scaling. The DA2 baseline predictions (right, red) show noticeable depth distortions, confirming that raw relative depth predictions alone fail to recover accurate metric depth. In contrast, the learned adaptive scaling (middle, blue) closely aligns with the oracle-scaled reconstruction (left, green), reducing scale errors and improving alignment with the ground truth depth structure.

These results reinforce our quantitative findings, showing that learning an image-conditioned scaling factor significantly improves metric depth recovery.



Oracle Scaling (Green) + Ground Truth PCD | Predicted Scaling (Blue) + Ground Truth PCD | Unscaled DA2 (Red) + Ground Truth PCD

Fig. 3: Qualitative evaluation via point cloud visualization. All three images contain the ground truth depth-based point cloud (RGB-colored). The left image overlays the oracle-scaled DA2 point cloud (green), the middle image overlays the learned adaptive-scaled DA2 point cloud (blue), and the right image overlays the unscaled DA2 point cloud (red), highlighting the significant misalignment of the baseline approach.

V. Conclusion

In this work, we explored the problem of monocular metric depth estimation and proposed a learned adaptive scaling approach to correct the scale ambiguity in relative depth predictions. Our method predicts an image-conditioned scaling factor, allowing models such as Depth Anything V2 (DA2) to produce accurate metric depth without requiring ground truth supervision during inference. Through extensive evaluations, we demonstrated that our learned scaling significantly improves depth accuracy, achieving performance close to the oracle-scaled baseline. In particular, our approach effectively reduces scale drift, leading to more geometrically consistent 3D reconstructions.

Our qualitative and quantitative analyses indicate that applying an appropriate scaling factor is crucial for recovering metric depth. The DA2 baseline, which lacks scale correction, exhibited severe depth distortions, whereas our adaptive scaling approach closely aligned with the ground truth. This suggests that incorporating image-conditioned priors can help bridge the gap between relative and absolute depth estimation. We also observed that oracle scaling provided only marginal improvements over our learned scaling, further validating the effectiveness of our approach.

VI. Future Work

While our method has demonstrated strong performance in monocular metric depth estimation, several directions remain for future exploration. One key observation is that some depth predictions do not merely suffer from a scale misalignment but instead exhibit entirely incorrect depth distributions. A promising direction is to develop a mechanism to estimate whether a given depth prediction is likely to result in a completely erroneous distribution before running inference. This could enable selective depth estimation or post-processing, ensuring that only reliable depth maps are used for downstream tasks.

Another important extension is to explore pixel-wise scaling instead of applying a single global scaling factor per image. Some regions within an image may require different scaling adjustments due to varying scene properties, such as depth discontinuities and occlusions. A pixel-adaptive scaling model could further enhance accuracy, especially in complex environments.

By addressing these challenges, we aim to develop a more adaptive and reliable monocular depth estimation approach, ensuring its applicability to real-world robotics, AR/VR, and autonomous navigation systems.

References

[1] R. Ranftl, P. Bojanowski, M. Caron, A. Joulin, J. Anderegg, A. Bochkovskiy, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *NeurIPS*, 2022.

[2] W. Yin, Y. Liu, C. Shen, and W. Tian, "Geometric structure preserving depth estimation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[3] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *NeurIPS*, 2014.

[4] J.-W. Bian, Z. Li, N. Wang, W.-Y. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," *NeurIPS*, 2019.

[5] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," *ICCV*, 2019.

[6] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Midas v3: A model for monocular depth estimation trained on 10 datasets," *arXiv preprint arXiv:2304.05826*, 2023.

[7] R. Wang, S. Wang, H. Chen, and J. Jia, "Lidar-augmented monocular depth estimation via geometry-aware attention mechanism," *CVPR*, 2021.

[8] B. Ummenhofer, H. Zhou, L. Kugler, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," *CVPR*, 2017.

[9] J. Watson, F. Aleotti, N. Buch, M. Firman, G. J. Brostow, and D. Turmukhambetov, "Temporal self-supervision for monocular 3d object detection," *NeurIPS*, 2021.

[10] X. Li, Y. Huang, L. Zhang, and J. Han, "Depth anything v2: Unified relative and metric depth estimation," *arXiv preprint arXiv:2402.01799*, 2024.

[11] T. Zhou, M. Brown, N. Snavely, and J. Malik, "Unsupervised learning of depth and ego-motion from video," *CVPR*, 2017.

[12] Y. Zhou, W. Yin, and C. Shen, "Adaptive scaling for depth estimation with motion cues," *CVPR*, 2023.

[13] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," *ECCV*, 2012.

[14] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2002–2011.

[15] L. N. Wasserstein, *Markov processes over denumerable products of spaces, describing large systems of automata*. Nauka, 1969.

[16] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.