Toward Robust Depth Fusion for Mobile AR With Depth from Focus and Single-Image Priors

Ashkan Ganj[*]
Worcester Polytechnic Institute

Hang Su[†]
Nvidia Research

Tian Guo[‡]
Worcester Polytechnic Institute

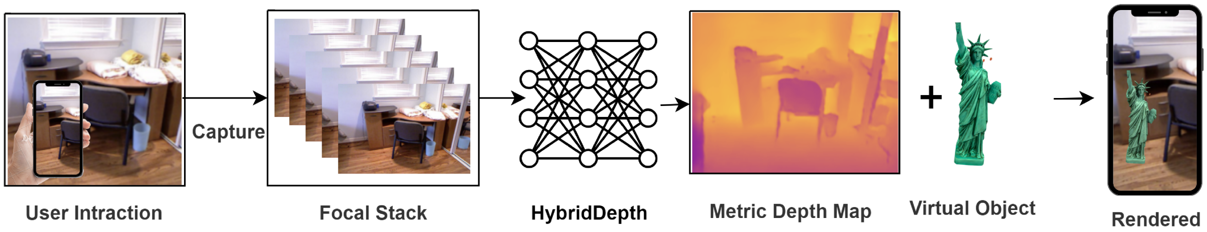

Figure 1: Example augmented reality workflow using HybridDepth. We can estimate accurate and robust dense metric depth from focal stacks captured by mobile devices for visually coherent rendering.

Abstract

We propose HybridDepth, a robust depth estimation pipeline that addresses the unique challenges of depth estimation for mobile AR, such as scale ambiguity, hardware heterogeneity, and generalizability. HybridDepth leverages the camera features available on mobile devices. It effectively combines the scale accuracy inherent in Depth from Focus (DFF) methods with the generalization capabilities enabled by strong single-image depth priors. By utilizing the focal planes of a mobile camera, our approach accurately captures depth values from focused pixels and applies these values to compute scale and shift parameters for transforming relative depths into metric depths. Through comprehensive quantitative and qualitative analyses, we demonstrate that HybridDepth not only outperforms state-of-the-art (SOTA) models in common datasets (DDFF12, NYU Depth v2). The source code of this project is available at https://github.com/cake-lab/HybridDepth.

Index Terms: Metric Depth Estimation, Augmented Reality, Depth From Focus, Depth Estimation

1 Introduction

Depth estimation is crucial for rendering visually coherent virtual scenes in augmented reality applications. For instance, in an online furniture shopping app, accurately placing a virtual chair next to a physical table requires a precise understanding of their relative distances from the mobile device to manage occlusions and to correctly size the chair based on its absolute distance [20]. Figure 1 shows an overview of the usage of metric depth maps in an object placement task.

Depth estimation techniques for mobile AR can be broadly divided into two categories: those that rely on specialized hardware such as light-field cameras [2], LiDAR sensors, or time-of-flight cameras [20], and those that utilize only the camera, e.g., monocular depth estimation. While specialized hardware can provide absolute metric depth information, this information is often sparse, and its availability is not guaranteed for all mobile devices. In contrast, monocular depth estimation approaches can predict depth maps even from a single image [3, 6]. While single-image depth estimation offers significant deployment advantages, enabling easy integration into various AR scenarios, it inherently suffers from scale ambiguity and generalization issues. These problems become particularly noticeable when such models are applied to real-world settings. Notably, prior work [7] showed that SOTA models like Zoedepth [3] fall short of maintaining performance consistency across diverse real-world environments.

Monocular depth estimation techniques are further categorized into metric depth [3] and relative depth methods [15]. Metric depth models, often trained on a limited number of datasets, tend to overfit and perform poorly in unseen environments or depth ranges. On the other hand, relative depth models excel in generalization and output more geometrically accurate depth maps; they are typically easier to train on larger and more varied datasets because they focus solely on the spatial relationships between elements within an image, removing the scale factor from the equation. However, they do not provide metric depth information, which is critical in AR applications where precise physical measurements are essential.

In this paper, we investigate the problem of providing robust metric monocular depth estimation for real-world scenarios. In order to achieve this, we need to solve two major challenges: 1) the scale ambiguity problem and 2) generalization. We propose a combined single-image relative and metric depth solution, called HybridDepth, to address both challenges. We choose to leverage the single-image priors from the relative depth estimation models because they are very good at generalizing across diverse environments and capturing essential geometrical details. Furthermore, we explore the depth from the focal stack (DFF) methods because they are very good at capturing metric depth but often suffer from poor generalization. To combine the strengths of both approaches, our work introduces a novel integration strategy that merges the reliable generalization and structural accuracy of relative depth models with the metric precision of DFF methods. HybridDepth is designed to deliver excellent zero-shot performance, effectively generalizing to unseen data or scenes, by using a three-stage approach: (1) *Capture results of metric and relative branches* (2) *Global scaling with Least-Squares Fitting* (3) *Scale Refinement Layer on globally scaled depth map.*

We have conducted comprehensive experiments to evaluate our method on well-known datasets such as NYU Depth v2, DDFF12. Our results demonstrate that our pipeline HybridDepth outperforms SOTA methods, including the recent DepthAnything work [19]. Specifically, we achieve a 13% improvement on average of the RMSE and AbsRel metrics on the NYU Depth v2 dataset.

In summary, our main contributions are:

[*] e-mail: aganj@wpi.edu
[†] e-mail:hangsu@nvidia.com
[‡] e-mail: tian@wpi.edu

- We design and implement an end-to-end pipeline HybridDepth that demonstrates the feasibility and potential of fusing focal stack information with relative depth to achieve robust metric depth estimation.
- HybridDepth establishes new SOTA results on two tested datasets, i.e., NYU Depth v2, DDFF12.

2 Related Work

Relative depth estimation focuses on the relative ordering of pixel depths without providing metric information, simplifying training and enhancing generalization. Recent Models [14, 13], achieve strong zero-shot performance through scale-invariant loss functions and diverse dataset training. These models excel in structural and segmentation accuracy, essential for AR tasks. Our work builds on these advancements to achieve robust metric depth performance.

Single-Image Metric Depth Estimation provides exact depth values using a single image. Approaches like ZoeDepth [3], and AdaBin [6] treat metric regression as a classification problem or leverage relative depth models on single RGB images. However, their performance on unseen data is limited [7].

Depth from Focus (DFF) estimates depth by identifying the sharpest focus distance for each pixel, utilizing the each pixel's sharpness level. Traditional methods [12] require large focal stacks, which are impractical due to data and time constraints. New deep learning methods [9, 1] address this by efficiently finding the best focal plane for each pixel but face challenges when suitable focal planes are missing, leading to potential inaccuracies.

3 HybridDepth: Robust Metric Depth Estimation

Problem Formulation. The goal is to estimate metric depth using only the camera and its features, determining the 3D metric information (distance from each pixel to the camera lens). This is a challenging task due to information loss during image capture. Our method must generalize across diverse environments and provide accurate geometric depth maps for mobile AR.

Solution. Figure 2 illustrates HybridDepth, a depth estimation pipeline leveraging Depth from Focus (DFF) and relative depth estimation. Inspired by prior works [17, 1], HybridDepth with novel modifications in *loss functions* (§4.1.1), *training processes* (§4.1), and *intermediate data processing* (§3.3) to optimize for focal stack processing. HybridDepth consists of three main stages. It begins by processing a focal stack, from which we select a single frame as the input for the relative depth branch and feed the entire stack into the DFF branch. The output from the relative depth branch forms the foundation of our depth map. Unlike other approaches [3], our pipeline focuses on preserving the structural accuracy and generalizability of relative depth models by scaling the relative depth into metric depth mathematically and refining it's scale errors with a small scale refinement model. The modular design of HybridDepth enhances flexibility, allowing each component (i.e., the DFF and the relative depth models) to be independently updated or replaced, thereby continuously improving performance.

3.1 Capturing Relative and Metric Depth ①

The first phase of our approach involves two key modules selected to generate the necessary intermediate data for the entire depth estimation pipeline: the Single-Image Relative Depth Estimator and the DFF Metric Depth Estimator. In the subsequent stages, we will use the metric information provided by the DFF module and fuse it with the relative depth map generated by the Single-Image module.

Single-Image Relative Depth Estimator. This module generates a relative depth map, which serves as the foundational layer for our depth estimation process. By using this depth map as a base, we ensure that the final output maintains structural integrity, producing sharp, well-defined edges and preserving object boundaries. We used a small version of Depth anything [19] for this module.

DFF Metric Depth Estimator. This module provides the critical scale and metric information necessary to convert the relative depth map into a metric depth map. Given a focal stack as input, the DFF module produces a dense metric depth map of the scene. This metric depth map is then used to convert the relative depth map from the Single-Image Relative Depth Estimator to the metric depth map. In implementation we used the Depth from Focal Stack model DFV [1], specifically employing differential focus volumes to capture depth map.

3.2 Fusing Relative and Metric Depth information ②

This is the first step we are trying to mathematically fuse metric information from DFF branch to the relative depth.

The Global Scale and Shift alignment. The global scaler transforms relative depth data into metric depth using scale and shift adjustments, as modeled by Equation (1):

$$\text{Metric Depth} = \text{Scale} \times \text{Relative Depth} + \text{Shift} \tag{1}$$

The Scale and Shift parameters are calculated using the least-squares fitting technique with the Metric Depth value from DFF. This approach aligns the relative depth map with the DFF metric data, minimizing discrepancies and producing an intermediate metric depth output. This method is different from other conventional methods [3, 19] that use deep learning-based models to fuse two depths and can preserve the integrity of the original relative depth map. This method combines the strengths of relative depth estimation with precise metric information.

3.3 Refinement. ③

Global scale and shift alignment can introduce errors, as it attempts to convert the entire relative depth map to metric depth using just two numbers. This simplification can lead to inaccuracies in some pixels and regions. To address this, we first calculate the scale difference between the globally scaled depth map from step ② and the DFF branch output from step ①. This allows us to build a new scale map by dividing these two depth maps. However, since both the DFF branch and the globally scaled depth maps can contain errors, the resulting scale map may also be imprecise. Consequently, we introduce a refinement layer that applies local scale corrections to different pixels of the globally scaled depth map using the scale map derived from DFF.

The Scale Refinement Layer. To construct this layer, we utilize a customized version of MiDaS-small [15] initialized with pretrained ImageNet [4] weights, similar to the model used in [17], to correct scale errors in individual pixels. Our refinement approach differs from that of [17], which deals with sparse depth and uses scale regression to fill empty regions of the scale map (comes from sparse depth). Instead, we leverage all the depth values from the DFF to build a scale map based on the globally scaled depth map. This method allows us to effectively apply scale refinement to each pixel.We feed it an input of two concatenated data channels: the globally scaled depth map and the DFF-derived scale map. This approach allows the scale refinement model to learn and apply local scale adjustments, enhancing the overall accuracy.

4 Implementation Details

4.1 Training

We use the AdamW optimizer, configured with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\lambda = 0.001$. Training involves different learning rates adjusted according to the dataset: 3×10^{-4} for the NYU Depth dataset, 1×10^{-5} for DDFF12. We use one NVIDIA A100 40GB GPU for training, with dataset-specific batch size: 24 for NYU Depth v2, 8 for DDFF12. We trained the model until the validation loss converges and pick the one with lowest loss.

We use the original data size for NYU Depth v2 and for DDFF12, the input size is set to 224×224 pixels with random crop and flip augmentations applied for training but used the original image size of 383×552 for evaluation like other DFF-based methods [1]. Frames in focal stacks were arranged in ascending order of focal distance to maintain consistency of depth processing.

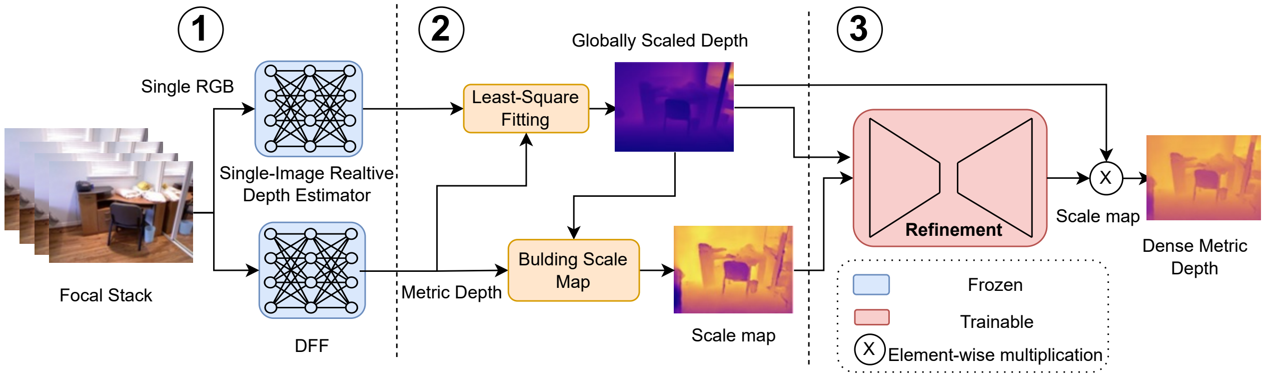

Figure 2: An overview of HYBRIDDEPTH which consists of three stages: (1) capture a focal stack and pass the frames through two branches; (2) calculate scale and shift based on estimated relative and metric depth maps using least-squares fitting; (3) input a globally scaled depth map and a processed version of the Metric DFF branch output to the refinement model to output the updated scale map, which will be applied to the globally scaled depth map to get the final depth map.

4.1.1 Loss Function

Previous work like VI-Depth [17] uses L1 loss for regression tasks, but it is sensitive to distance range changes, affecting zero-shot performance. To improve this, we adopt the *scale-invariant* loss function L_{SILog} from [5] and incorporate a multi-scale gradient loss function L_{grad} to enhance visual quality and boundary preservation. The overall loss function L is defined as:

$$L = L_{\text{SILog}} + 0.5 \times L_{\text{grad}}, \tag{2}$$

where L_{SILog} is:

$$L_{\text{SILog}} = 10 \times \sqrt{\text{var}(g) + \beta \times (\text{mean}(g))^2}, \tag{3}$$

with $g = \log(d + \alpha) - \log(d_{gt} + \alpha)$.
The gradient loss L_{grad} is:

$$L_{\text{grad}} = \frac{1}{HW} \sum_{s=1}^{4} \sum_{i,j} \left| \nabla_s d_{i,j} - \nabla_s d_{gt,i,j} \right| \tag{4}$$

Here, ∇ denotes the spatial gradient, s is the scale factor, d is the predicted depth, d_{gt} is the ground truth depth, and H and W are the depth map dimensions. This combined loss optimizes both scale-invariant and gradient-based aspects, enhancing depth map accuracy and geometric detail.

4.2 Data Synthesizing

The ability to synthesize focal stack is vital for overcoming the limitations associated with the availability of datasets containing real focal stacks. To develop a model capable of operating effectively across various AR scenarios and allowing robust comparisons with state-of-the-art models, we used a method from [9, 16] to artificially recreate focal stacks from a single image to train and evaluate our model on datasets like NYU Depth v2.

5 EXPERIMENTS

We evaluated HYBRIDDEPTH's performance on different datasets. Our method combines monocular depth estimation and depth-from-focus (DFF) solutions. So we compared HYBRIDDEPTH with both single-image and DFF models.

A challenge in comparing our method with SOTA models was the lack of a common benchmark for both DFF and monocular depth models that included focal stack images. We addressed this by selecting NYU Depth v2 as a single-image depth dataset and creating synthesized focal stacks using the method in §4.2. Additionally, we used a real-world DFF-based dataset to compare HYBRIDDEPTH with other models using focal stack images as input.

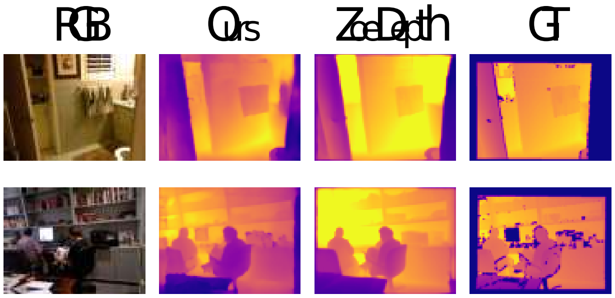

Figure 3: Qualitative comparisons between our model, and ZoeDepth [3] on the NYU Depth v2 dataset.

Table 1: Performance comparison on the NYU Depth v2 dataset with focal stack size of 10. **Bold** values represent the best results. The evaluation uses an upper bound of 10 meters on the ground truth depth map. All the numbers for other works have been taken from the corresponding papers.

Model	**Type***	**RMSE** ↓	**AbsRel** ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
AdaBin[6]	SIDE	0.364	0.103	0.903	0.984	0.997
ZoeDepth [3]†	SIDE	0.270	0.075	0.96	0.995	0.999
VPD [21]	SIDE	0.254	0.069	0.96	0.995	0.999
Depth Anything [19]	SIDE	0.206	0.056	0.984	**0.998**	**1.000**
Ours	DFF	**0.202**	**0.041**	**0.988**	**0.998**	**1.000**

* *SIDE* stands for single image depth estimation.
† We used ZoeDepth-M12-N version.

- **DDFF dataset** [8] is a real-world DFF dataset captured using a light-field camera across 12 scenes. The training set consists of six scenes with 100 samples each, and the test set includes six scenes with 20 samples each. Each sample comprises a 10-frame focal stack with a corresponding ground truth disparity map at a resolution of 383 × 552 pixels. For our experiments, we used a focal stack of 5 frames.
- **NYU** [10] is a benchmark for monocular depth estimation in indoor scenes. It includes over 24K labeled RGB and depth image pairs in the training set and 654 pairs in the test set, with ground truth depth maps at a resolution of 640 × 480 pixels.

Our experiments show that our model outperforms current SOTA models on all the mentioned datasets by a good margin. We also provide a qualitative comparison between our model's depth maps and a SOTA model [3].

5.1 Comparison to the State-of-the-Art

As discussed earlier, to ensure a fair comparison, we directly compared our model with other works on the specific datasets that they

Table 2: Performance comparison on the DDFF12 dataset. **Bold** values represent the best results. We used the same split as DFV [1]. All the numbers for other works have been taken from the DFV paper.

Model	MSE ↓	RMSE ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
DFV [1]	5.70×10^{-4}	0.0213	0.17	0.76	**0.94**	**0.98**
Defocus-Net [9]	8.61×10^{-4}	0.0255	0.17	0.61	**0.94**	0.97
DDFF [8]	8.97×10^{-4}	0.0276	0.24	0.61	0.88	0.96
DFFintheWild [18]	5.7×10^{-4}	-	0.17	0.776	0.874	0.939
Ours	$\mathbf{5.58 \times 10^{-4}}$	**0.0205**	**0.16**	**0.79**	**0.94**	**0.98**

trained on. We achieved this by using two different types of works, namely state-of-the-art (SOTA) single image depth models [3, 6, 21, 19] and models based on depth from focus/defocus [1, 9].

Results on DDFF12. This dataset is a challenging dataset for depth-from-focus (DFF) methods since it contains large texture-less areas where focus cues are not very visible in the focal stack. Table 2, shows that our model achieves excellent results and outperforms the current SOTA model [1] on this dataset, with an MSE of 5.6×10^{-4}, an RMSE of 0.0205, and an AbsRel of 0.16. These results demonstrate that our model can effectively address scale inaccuracies through an additional layer of scale refinement and perform better than DFF models, specifically in cases with texture-less regions and weak focus cues.

Results on NYU Depth v2. Table 1 shows that our model achieves a new state-of-the-art performance on this dataset and outperforms all the other models, including single image and DFF-based methods. Our model shows a smaller amount of error on all of our evaluation metrics. The models HYBRIDDEPTH outperforms including more complex models such as VPD [21], Depth Anything [19], ZoeDepth [3]). This result highlights the efficacy of using focal stack clues for depth estimation task. Figure 3 shows the qualitative comparison of our work with Zoedepth [3]. Our model demonstrates a better visual quality and outputs smoother and more accurate depth maps. Unlike Zoedepth, our model is also capable of capturing depth over long distances. Also, HYBRIDDEPTH can capture some small objects depth details.

6 Conclusion and Future Work

Achieving robust and accurate metric depth in the wild is a challenging problem. Recent work has demonstrated that even SOTA models like Zoedepth struggle with real-world AR scenarios. We tackle this challenge with the design of HYBRIDDEPTH, an end-to-end metric depth estimation pipeline that fuse the focal stack and relative depth information. We show that HYBRIDDEPTH outperforms both single image and DFF models on commonly used datasets: NYU Depth v2 and DDFF12.HYBRIDDEPTH's superior performance only requires the use of cameras, which are widely available on almost all mobile devices. Compared to solutions that rely on specialized hardware like LiDAR or ToF sensors, HYBRIDDEPTH is more deployment friendly. Currently, the DFF branch in HYBRIDDEPTH represents the most significant source of errors in our pipeline, particularly due to scaling errors in situations where the focal stack does not include an ideal focus for certain pixels. As part of future work, we will investigate methods to selectively capture depth values that are close to the focus distance, thereby ensuring the accuracy of the provided depth values.

Acknowledgments

This work was supported in part by NSF Grants #2105564, #2236987, and #2346133, a VMware grant, the Worcester Polytechnic Institute's Computer Science Department. Most results presented in this paper were obtained using CloudBank [11], supported by the National Science Foundation under award #1925001.

References

[1] Deep depth from focus with differential focus volume. In *Proceedings - 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2022. 2, 4

[2] G. Baruch, Z. Chen, A. Dehghan, T. Dimry, Y. Feigin, P. Fu, T. Gebauer, B. Joffe, D. Kurz, A. Schwartz, and E. Shulman. ARKitScenes - A Diverse Real-World Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data. In *NeurIPS Datasets and Benchmarks Track*, 2021. 1

[3] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv:2302.12288*. doi: 10.48550/ARXIV.2302.12288 1, 2, 3, 4

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848 2

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

[6] S. Farooq Bhat, I. Alhashim, and P. Wonka. AdaBins: Depth Estimation Using Adaptive Bins. In *CVPR*, 2021. 1, 2, 3, 4

[7] A. Ganj, Y. Zhao, H. Su, and T. Guo. Mobile AR Depth Estimation: Challenges & Prospects. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, HOTMOBILE '24. Association for Computing Machinery, New York, NY, USA, 2024. 1, 2

[8] C. Hazirbas, S. G. Soyer, M. C. Staab, L. Leal-Taixé, and D. Cremers. Deep depth from focus. In *Asian Conference on Computer Vision (ACCV)*, December 2018. 3, 4

[9] M. Maximov, K. Galim, and L. Leal-Taixe. Focus on defocus: Bridging the synthetic to real domain gap for depth estimation. In *CVPR*, 2020. 2, 3, 4

[10] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 3

[11] M. Norman, V. Kellen, S. Smallen, B. DeMeulle, S. Strande, E. Lazowska, N. Alterman, R. Fatland, S. Stone, A. Tan, K. Yelick, E. Van Dusen, and J. Mitchell. Cloudbank: Managed services to simplify cloud access for computer science research and education. In *Practice and Experience in Advanced Research Computing*, PEARC '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3437359.3465586 4

[12] S. Pertuz, D. Puig, and M. A. Garcia. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5):1415–1432, 2013. doi: 10.1016/j.patcog.2012.11.011 2

[13] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. doi: 10.1109/ICCV48922.2021.01196 2

[14] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 2

[15] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2

[16] H. Si, B. Zhao, D. Wang, Y. Gao, M. Chen, Z. Wang, and X. Li. Fully self-supervised depth estimation from defocus clue. *arXiv preprint arXiv:2303.10752*, 2023. 3

[17] Wofk, Diana and Ranftl, René and Müller, Matthias and Koltun, Vladlen. Monocular Visual-Inertial Depth Estimation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 2, 3

[18] C. Won and H.-G. Jeon. Learning depth from focus in the wild, 2022. 4

[19] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 1, 2, 3, 4

[20] Y. Zhang, T. Scargill, A. Vaishnav, G. Premsankar, M. Di Francesco, and M. Gorlatova. Indepth: Real-time depth inpainting for mobile augmented reality. *IMWUT*, 2022. 1

[21] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu. Unleashing text-to-image diffusion models for visual perception. *ICCV*, 2023. 3, 4