arXiv:2407.18443v3 [cs.CV] 25 Dec 2024

HybridDepth: Robust Metric Depth Fusion by Leveraging Depth from Focus and Single-Image Priors

Ashkan Ganj
Worcester Polytechnic Institute
aganj@wpi.edu

Hang Su
Nvidia Research
hangsu@nvidia.com

Tian Guo
Worcester Polytechnic Institute
tian@wpi.edu

Abstract

We propose HYBRIDDEPTH*, a robust depth estimation pipeline that addresses key challenges in depth estimation, including scale ambiguity, hardware heterogeneity, and generalizability.* HYBRIDDEPTH *leverages focal stack, data conveniently accessible in common mobile devices, to produce accurate metric depth maps. By incorporating depth priors afforded by recent advances in single-image depth estimation, our model achieves a higher level of structural detail compared to existing methods. We test our pipeline as an end-to-end system, with a newly developed mobile client to capture focal stacks, which are then sent to a GPU-powered server for depth estimation.*

Comprehensive quantitative and qualitative analyses demonstrate that HYBRIDDEPTH *outperforms state-of-the-art (SOTA) models on common datasets such as DDFF12 and NYU Depth V2.* HYBRIDDEPTH *also shows strong zero-shot generalization. When trained on NYU Depth V2,* HYBRIDDEPTH *surpasses SOTA models in zero-shot performance on ARKitScenes and delivers more structurally accurate depth maps on Mobile Depth. The code is available at https://github.com/cake-lab/HybridDepth/.*

1. Introduction

Depth estimation is a critical task in computer vision, with applications [14] spanning autonomous driving, augmented reality [10, 13], and robotics [31]. While hardware like LiDAR and Time-of-Flight (ToF) sensors are commonly used for more expensive applications, they are often not available for mobile and consumer-specific applications. This has led to extensive research on vision-based depth estimation methods that rely solely on cameras. Monocular depth estimation models have gained popularity due to their simplicity and minimal hardware requirements; however, these models frequently suffer from scale ambiguity, where depth estimations vary across different zoom levels. Additionally, single-image models, such as ZoeDepth [4] and ECoDepth [18], struggle to generalize well to real-world conditions, as demonstrated by recent work [9] using ARKitScenes [2].

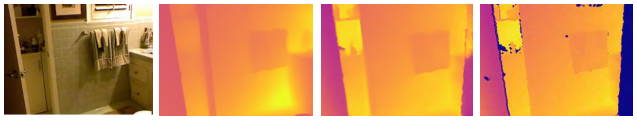

Figure 1. HYBRIDDEPTH produces globally scaled depth maps, and refines them to further correct errors and enhance details.

Depth-from-focus (DFF) methods, which leverage focal stack information, provide a promising alternative by offering more robust depth cues and generating reliable metric depth estimates using only cameras. However, existing DFF-based models like DFV [28] tend to produce noisy estimations in texture-less or challenging regions and can't generalize well, limiting their overall effectiveness. On the other hand, relative depth models, while capable of generalizing well to unseen data and capturing fine details in depth maps, do not provide metric information, making them unsuitable for applications that require metric depth values.

In this work, we address the above mentioned limitations by combining DFF with relative depth models. While relative depth models excel in generalization and structural accuracy, DFF provides reliable metric depth information without requiring specialized hardware. By integrating these two approaches, we have the potential to leverage the strengths of both, delivering robust and accurate depth estimation that is both scalable and applicable to a wide range of real-world scenarios.

Intuitively simple, we have to address the challenge of effectively combining focal stack information with relative depth models to maintain the *structural accuracy and generalization capabilities* of relative depth estimation, while incorporating the *reliable metric depth information* from DFF. Toward this end, we propose an end-to-end solution, called HYBRIDDEPTH, that fuses DFF and single-image priors to achieve robust metric depth estimation.

HYBRIDDEPTH is designed to deliver zero-shot performance, *i.e.* to effectively generalize to unseen data or scenes, by employing a three-stage approach. First, we

capture the outputs from both the relative depth and metric depth models. Next, we perform a global scaling that aligns the relative depth output with the absolute scales provided by the metric information from DFF. Finally, we apply a deep learning-based refinement layer to fine-tune the intermediate depth map (i.e., globally scaled depth map), smoothing out any inconsistencies and enhancing the overall accuracy of the depth estimations. Figure 1 provides an example of how HYBRIDDEPTH refines the globally scaled depth map.

We conduct comprehensive experiments to evaluate our method on two real-world focal stack datasets, DDFF12 and Mobile Depth[1]. Due to the lack of focal stack datasets, we evaluate HYBRIDDEPTH on additional datasets, including NYU Depth V2 and ARKitScenes, with synthesized focal stacks. Our results demonstrate that HYBRIDDEPTH establishes new SOTA on depth estimation with focal stacks, as well as improved generalization capabilities. The qualitative zero-shot results on Mobile Depth and ARKitScenes show that HYBRIDDEPTH also generalizes well.

For example, HYBRIDDEPTH achieves a 10.5% and 6.1% improvement of MSE and RMSE on DDFF12, and 6.5% and 7.1% improvement of RMSE and AbsRel on the NYU Depth V2 dataset, compared to DFV [28]. Moreover, our zero-shot performance on the ARKitScenes dataset shows a 43% improvement compared to SOTA methods [29]. We also conduct comparisons with single-image depth estimation approaches such as Depth Anything [29] to demonstrate the significant advantage of utilizing the focal stack information on mobile devices.

In summary, our main contributions are as follows.

- We design and implement an end-to-end pipeline HYBRIDDEPTH that demonstrates the feasibility and benefits of fusing focal stack information with single-image priors in achieving robust metric depth estimation. *Relevant research artifacts will be open-sourced.*
- HYBRIDDEPTH establishes new SOTA on DFF-based depth estimation, with a compact model of 240 MB and an inference time of 20 ms, which is 19.2% of the Depth Anything size and 2.85X faster.
- HYBRIDDEPTH showcases strong generalization performance on Mobile Depth and the AR-specific dataset ARKitScenes.
- We demonstrate the significant advantage of HYBRIDDEPTH over single-image depth estimation approaches, offering a viable and attractive alternative for mobile applications.

2. Related Work

Single-image depth estimation outputs depth maps given only single input images. Due to the inherent ill-posedness of the task, such approaches usually adopt an affine-invariant depth map formulation. This specific formulation enables training on large datasets, and has been proven to be key to model generalization. For instance, models like Midas [5, 20], DPT [19], and Depth Anything [29] have made a good progress in zero-shot performance by using a novel scale-invariant loss function, enabling training on datasets captured with various hardware devices, and new way of using data from different domains. Also, these models are very good at maintaining structural and segmentation accuracy. Our work leverages the recent advancements in relative depth in zero-shot performance [29] as the basis for achieving robust metric depth performance.

Metric depth estimation aims to provide exact depth values in physical units, such as meters. While this was traditionally only feasible with stereo or video approaches, models like Depth Anything [29], ZoeDepth [4], AdaBin [8], and LocalBin [3] have shown success with single image inputs. However, such approaches usually degrade significantly for unseen data [9] and unknown camera models [30]. In contrast, our work relies less on visual data and, therefore, can circumvent the current problems of metric depth estimation.

Depth from focus (DFF) estimates depth by identifying the focus distance at which each pixel is most sharply defined, while areas outside the focal plane appear blurry, creating a circle of confusion (CoC). Deep learning-based methods [12, 15, 24, 28] use convolutional neural networks (CNNs) and neural networks to find the best focal plane for each pixel from a limited number of images. However, DFF can be noisy when suitable focal planes are missing from the data or when dealing with texture-less regions. Our design effectively mitigates these limitations by combining DFF with relative depth information and refining the depth map, ensuring accurate and stable estimations.

3. HYBRIDDEPTH

We propose HYBRIDDEPTH, a three-stage pipeline that achieves generalizable metric depth without subjecting to the scale ambiguity issue. The design of HYBRIDDEPTH is inspired by ViDepth [25], which uses IMU sensors. Unlike ViDepth, HYBRIDDEPTH relies only on RGB inputs, fulfilling the goal of providing highly detailed metric depth maps without the burden of additional sensors on mobile applications. Our key idea is to *leverage a well-generalized relative depth model as the basis and then convert the relative depth to metric depth with the help of a DFF model*. The focal stack provides precise depth information using just a camera [28], and relative depth estimation is known for its ability to generalize well across different scenes while maintaining strong structural accuracy [20, 29]. By combining these two methods, we show that HYBRID-

[1]Only qualitative results are shown due to lack of ground truth depth.

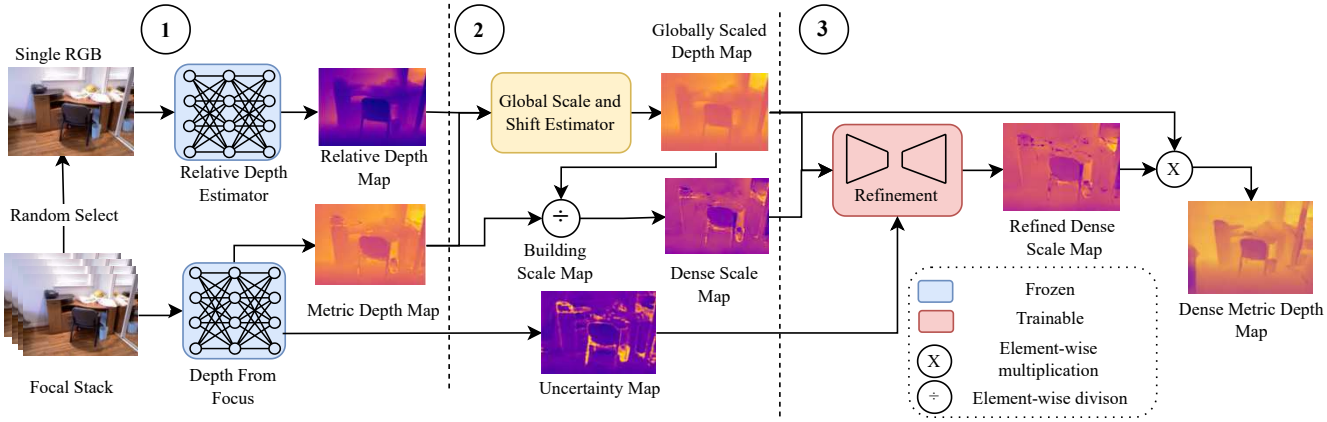

Figure 2. An overview of HYBRIDDEPTH which consists of three stages: (1) capture a focal stack and pass the frames through two branches; (2) calculate scale and shift based on estimated relative and metric depth maps using least-squares fitting; (3) input a globally scaled depth map and a processed version 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.

DEPTH can benefit from both worlds and outperform both methods [20, 28].

Figure 2 describes the overall architecture of HYBRIDDEPTH. In essence, the pipeline performs pixel-wise linear transformation to convert each pixel's relative depth to metric depth. The pipeline begins by processing the input focal stack. A randomly selected frame within the focal stack is passed through the relative depth estimator to generate a relative depth map, while the entire focal stack is fed into the DFF branch to produce a metric depth map (1). These outputs are then aligned using a global scale and shift estimation process (2). Finally, we address the remaining scale errors that occurred in (2) with a trainable scale refinement network (3). In the following sections, we provide details about each of the three modules within our pipeline.

3.1. Capturing Relative and Metric Depth (1)

The first step of HYBRIDDEPTH consists of two key modules, a *single-image relative depth estimator* and a *DFF metric depth estimator*, to generate the intermediate data for the later metric conversion and refinement. Note that even though the DFF module can output the metric depth map, the depth map quality can be low and lack details, as we will show later in the evaluation.

Relative Depth Estimator. This module generates a dense relative depth map, serving as the foundation for our depth estimation process. It is designed to utilize a pre-trained model that takes a single RGB image as an input and produces a dense relative depth map. The reason we use a pre-trained model like Depth Anything[2] [29] is because it is trained on large amounts of datasets and, therefore, has a potential for good zero-shot performance.

DFF Metric Depth Estimator. This module provides the crucial metric information needed to convert the relative depth map into a generalizable metric depth map. Unlike other methods that rely on external sensors [32], our approach is entirely visual-based. By leveraging focus cues from a *focal stack*, consisting of images captured at different focus distances, this model assigns each pixel to its corresponding focus distance, yielding the metric depth for that specific pixel. For this module, we employ DFV[3] [28], a lightweight SOTA model that produces accurate metric depth for the DDFF12 dataset.

3.2. Relative and Metric Depth Fusion (2)

The Global Scale and Shift Alignment. This module performs the linear transformation of the relative depth map to a globally scaled depth map using Equation (1).

$$\text{Metric Depth} = \text{Scale} \times \text{Relative Depth} + \text{Shift} \quad (1)$$

Here, the *Scale* and *Shift* parameters are obtained via the least-square fitting [20] using the metric and relative depth maps obtained from (1). Applying the scale value globally for all pixels brings relative depth values to a correct order of magnitude, while applying global shift can help undo potential bias or offset in the original estimation. Unlike prior work [4, 29] that often alter the fundamental depth relationships, scaling relative depth directly allows us to maintain such relationships, and therefore avoid distortions and ensure high-quality depth maps.

Building Dense Scale Map. To construct the dense scale map, we compute the pixel-wise scale difference between the globally scaled depth map (produced in (2)) and the DFF

[2]Note that Depth Anything can also output metric depth directly, which we outperform as will be shown in the evaluation section.

[3]Directly using models like DFV will not give us the generality as we will show in the evaluation.

branch's depth output (from ①). This process involves a straightforward pixel-level division, where each pixel in the globally scaled depth map is divided by the corresponding pixel in the DFF depth map. The resulting scale map captures the local scale variations, which will later be used in the refinement layer ③ to generate a refined dense scale map. This pixel-wise division ensures that the refinement layer can adjust each pixel's depth value according to its specific scale discrepancy.

3.3. Scale Refinement Layer ③

Global scale and shift alignment can introduce errors, as it attempts to convert the entire relative depth map into metric depth using only two parameters. This oversimplification may result in inaccuracies across specific pixels and regions. Our experiments suggest that certain areas of the globally scaled depth maps can benefit from localized scale refinements. To address this, we introduce a refinement layer that applies pixel-wise scale corrections to the globally scaled depth map, utilizing the scale map and uncertainty map derived from the DFF module.

Our refinement process leverages a customized version of MiDaS-small [21] to correct pixel-wise scale errors. Specifically, this refinement model leverages the globally scaled depth map, the DFF-derived scale map, and the uncertainty map and outputs a refined dense scale map, which consists of pixel-wise depth scale adjustments. The uncertainty map allows the refinement layer to account for uncertainties from DFF, effectively guiding the model in weighting the influence of each pixel in scale refinement. This addition of uncertainty awareness empowers the refinement model to make more informed, precise adjustments, significantly improving the final depth map's accuracy and robustness. By integrating visual cues, pixel-wise scale refinement, and uncertainty-driven adjustments, this layer offers a novel and highly effective approach to refining depth maps, delivering enhanced metric precision without reliance on external sensors.

4. Implementation Details

Our model is implemented using the PyTorch framework. For real-world testing, we developed a mobile client using the Android Camera2 API to capture focal stacks, coupled with an edge server equipped with an NVIDIA RTX 4090 GPU for inference.

4.1. Training

We use Depth Anything pre-trained weights in all experiments to leverage its generalization. For the DFV module, we either use pre-trained weights (DDFF12, Mobile Depth, and ARKitScenes) or train it from scratch (NYU Depth V2). For the zero-shot experiments on ARKitScenes, we use both the pre-trained DFV models from NYU Depth V2 and DDFF12, with the refinement layer trained on NYU Depth V2. For Mobile Depth, we use the pre-trained DFV and train the refinement layer on NYU Depth V2.

All models are trained using the AdamW optimizer, with dataset-specific hyperparameters and batch sizes. Further training details, including hyperparameters, input sizes, and augmentations for each dataset, can be found in Supplementary Material.

4.1.1 Loss Function

Prior work like ViDepth [25] utilizes L1 loss as their regression task loss function. However, we know that L1 loss is sensitive to changes in distance ranges, hindering performance on unseen data [7]. To address this issue, we adopt the *scale-invariant* loss function L_{SILog} proposed in [7]. Additionally, we integrate a multi-scale gradient loss function L_{grad} to enhance visual quality and sharpness while preserving image boundaries as much as possible. Our overall loss function L is as follows:

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

where L_{SILog} is defined as:

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

Here, $g = \log(d + \alpha) - \log(d_{gt} + \alpha)$, with α being a small constant to prevent undefined logarithmic operations, and β serving as a scaling factor for the mean squared term. α and β are set to $1e^{-7}$ and 0.15.

L_{grad} is defined as:

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

where ∇ denotes the first-order spatial gradient operator, s indicates the scale factor for multi-scale analysis, d represents the predicted depth map, and d_{gt} is the ground truth depth map. H and W are the height and width of the depth map, respectively. This composite loss function aims to optimize both the scale-invariant and gradient-based aspects of the predicted depth map, enhancing accuracy and geometrical information.

4.2. Data Synthesizing

The ability to synthesize focal stacks is crucial for overcoming the limitations of datasets that lack real focal stacks. To enable robust comparisons with SOTA models and to develop a versatile model for various applications (e.g., AR), we adopt a method to artificially recreate focal stacks from a single image with ground truth depth, similar to [22]. The process to artificially recreate focal stacks from a single image (with ground truth depth information) follows these

steps: (1) **Build an arbitrary camera system:** Configure a virtual camera with adjustable focus settings to mimic a physical camera system. (2) **Define focus distances:** Set specific focus distances to simulate camera focusing at different depths, similar to real-world camera behavior. (3) **Apply circular kernel for blurring:** Iterate over the image with a circular kernel to add blur based on the ground truth (GT) depth and the defined focus distances.

For the blurring process, we use Equation (5), which is the same equation that has been used in recent works [15, 22] for creating the synthesized defocus blur. This equation is used to determine the extent of blur for pixels outside the specific focal plane according to the GT depth.

$$c = \frac{|S_2 - S_1|}{S_2} \frac{f^2}{N \times (S_1 - f)}, \tag{5}$$

where f is the camera's focal length, N is the f-number (aperture) of the lens, S_1 is the distance to the in-focus subject, and S_2 is the distance beyond which subjects are considered out of focus. This equation allows for realistic synthesis of focal stacks by adjusting the blur based on depth, transforming a single image into multiple focal stack images. This enables us to leverage single-image depth datasets like NYU Depth V2 for training.

4.3. End-To-End Mobile Pipeline

We designed a mobile pipeline for utilizing HYBRIDDEPTH for depth estimation on a mobile client. The process begins with the mobile client capturing a focal stack of images(5 or 10), each representing different focus distances of the same scene. We developed an Android app using the Camera2 API to efficiently capture these focal stacks, adjusting the focus plane across five different values in a short time frame (approximately 141 ± 20 ms on a Pixel 6 Pro). All images are resized to 480×640 for uniform processing and reducing computational cost.

To address potential misalignment among focal stack images, we leverage the mobile device's built-in optical image stabilization sensor (OIS) during capture. This ensures the focal stack is properly aligned before being sent to our model pipeline. Once the images are captured, they are transmitted to a server equipped with an NVIDIA RTX 4090 GPU for processing. On the server, we first apply OCR-based image alignment using OpenCV, before feeding the focal stack to HYBRIDDEPTH. The resulting dense depth map, which can be utilized in various mobile applications such as AR, is then returned to the mobile device.

5. Experiments

A key challenge in evaluating HYBRIDDEPTH is the lack of commonly used datasets that can allow us to directly compare with single-image and DFF-based methods. To address this, we evaluated HYBRIDDEPTH's performance across four different datasets. For direct comparison with other DFF-based models, we utilized the DDFF12 dataset. Additionally, we synthesized the NYU Depth V2 dataset as described in Section 4.2 to facilitate comparisons between single-image methods and DFF-based approaches. To assess the generalizability of HYBRIDDEPTH, we conducted zero-shot evaluations on ARKitScenes (quantitatively, qualitatively), as well as Mobile Depth (qualitatively).

HYBRIDDEPTH consistently outperforms existing methods on all datasets. For example, HYBRIDDEPTH achieves a 6.1% and a 36% improvement in RMSE on the DDFF12 and NYU Depth V2 datasets compared to each dataset's specific SOTA model (Tables 1, 3, 2), respectively. Additionally, HYBRIDDEPTH demonstrates superior generalizability in zero-shot evaluations. For example, HYBRIDDEPTH achieves a 82.7% improvement in RMSE compared to Depth Anything (Table 4). Furthermore, our qualitative experiments (Figures 4, and 5) show that HYBRIDDEPTH produces higher-quality depth maps compared to other models. Please refer to Supplementary Material for more visualizations and results for inference time.

5.1. Datasets

We select a diverse set of datasets, including real-world and synthetic datasets, to comprehensively evaluate HYBRIDDEPTH and compare against single-image and DFF-based methods. The DDFF12 dataset offers real-world focal stacks captured by a light-field camera, enabling direct comparison with other DFF-based models. The Mobile Depth dataset, similar to DDFF12, contains real-world DFF data captured using a mobile phone and is used for qualitative comparison. We also transform the following two datasets with synthesized focal stacks to enable more comparisons. NYU Depth V2 is a widely recognized benchmark for monocular depth estimation, particularly for indoor scenes, which helps us compare HYBRIDDEPTH with other single-image depth estimation models. Lastly, ARKitScenes is a large-scale, diverse dataset captured with mobile devices that represents real-world challenges in depth estimation [9]. Please refer to Supplementary Material for more details about datasets.

5.2. Evaluation Metrics

We evaluate the metric depth accuracy with the following metrics: MSE as $\frac{1}{M}\sum_{i=1}^{M}(d_i - \hat{d}_i)^2$, RMSE as $\sqrt{\frac{1}{M}\sum_{i=1}^{M}(d_i - \hat{d}_i)^2}$, and AbsRel Error as $\frac{1}{M}\sum_{i=1}^{M}\left|\frac{d_i - \hat{d}_i}{d_i}\right|$. Here, d_i and $\hat{d}_i$ denote the ground truth and predicted depth at pixel i, and M is the total number of pixels in the image. Additionally, we assess the accuracy at threshold values using δ_1, δ_2, and δ_3 which measure the percentage of pixels where the predicted depth $\hat{d}_i$ is within 1.25, 1.25^2,

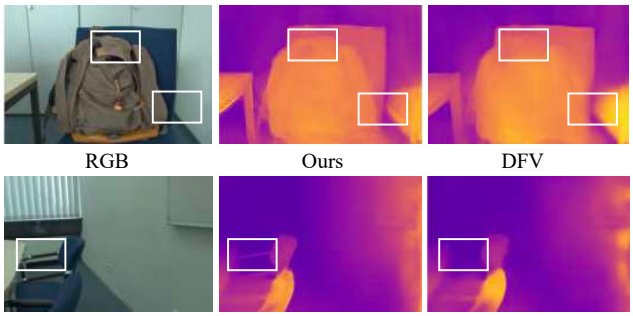

Figure 3. HYBRIDDEPTH performance in capturing small details in depth maps in comparison to DFV on DDFF12.

Table 1. Performance comparison on the DDFF12 dataset. **Bold** values represent the best results. All numbers for other works have been taken from the DFV paper. The unit for all metrics is disparity.

Model	MSE ↓	RMSE ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
RDF [16]	91.8×10^{-4}	0.0941	1.00	0.16	0.33	0.47
Defocus-Net [15]	8.6×10^{-4}	0.0255	0.17	0.61	0.94	0.97
DDFF [12]	8.9×10^{-4}	0.0276	0.24	0.61	0.88	0.96
DFFintheWild [26]	5.7×10^{-4}	-	0.17	0.78	0.87	0.94
DFV [28]	5.7×10^{-4}	0.0213	0.17	0.76	0.94	0.98
Ours	$\mathbf{5.1 \times 10^{-4}}$	**0.0200**	**0.17**	**0.79**	**0.95**	**0.98**

Table 2. Performance comparison on the NYU Depth V2 dataset with other Depth from Focus/Defocus methods. **Bold** values represent the best results. The evaluation uses an upper bound of 10 meters on the ground truth depth map. DefocuseNet [15] results have been taken from the corresponding paper. The unit for all metrics is disparity.

Model	Type*	RMSE ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
DefocusNet [15]	DFD	0.493	-	-	-	-
DefocusNet ($\leq 2m$)‡	DFD	0.180	-	-	-	-
DFV [28]	DFF	0.136	0.028	**0.996**	1.000	1.000
Ours	DFF	**0.128**	**0.026**	0.995	**1.000**	**1.000**
Ours ($\leq 2m$)‡	DFF	0.082	0.034	0.988	0.998	1.000

‡ These rows show performance metrics for distances under 2 meters.
* *DFD/DFF* stand for depth from defocus/focus depth estimation.

and 1.25^3 times the ground truth depth d_i, respectively. For the DDFF12 dataset, we use the same metrics for disparity calculation.

5.3. Comparison to the State-of-the-Art

Results on DDFF12. This dataset presents a significant challenge for depth-from-focus (DFF) methods due to large texture-less areas, where focus cues are often weak in the focal stack, leading to increased error possibilities. As shown in Table 1, our model outperforms the current state-of-the-art models, achieving a 10.5% improvement in MSE and a 6.1% improvement in RMSE compared to DFV [28]. Additionally, compared to Defocus-Net [15], our model achieves a 40.7% improvement in MSE and a 21.6% improvement in RMSE. These improvements highlight the effectiveness of our combination of Depth Anything and DFV with the refinement layer, which helps address scale inaccuracies and handle texture-less regions with weak focus cues. Qualitative results, as shown in Figure 3, further demonstrate that our model produces more detailed and higher-quality depth maps, preserving structural information from Depth Anything. This illustrates the strength of our pipeline in leveraging Depth Anything's depth cues and improving upon DFF models through scale refinement.

Table 3. Performance comparison on the NYU Depth V2 dataset with single-image depth estimation models. **Bold** and underlined values represent the best and second-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. The unit for all metrics is meter.

Model	Type*	RMSE ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
ZoeDepth [4]†	SIDE	0.270	0.075	0.96	0.995	0.999
VPD [33]	SIDE	0.254	0.069	0.96	0.995	0.999
ECoDepth [18]	SIDE	0.218	0.059	0.97	0.997	0.999
Depth Anything [29]	SIDE	<u>0.206</u>	<u>0.056</u>	<u>0.98</u>	<u>0.998</u>	**1.000**
Ours	DFF	**0.128**	**0.026**	**0.99**	**1.000**	**1.000**

* *SIDE* stand for single image depth estimation.
† For ZoeDepth we have used ZoeDepth-M12-N version.

Results on NYU Depth V2. Table 2 compares HYBRIDDEPTH with DFF-based methods. We trained the DFV module on the NYU Depth V2 dataset and used it as a baseline for comparison. With the addition of our refinement layer, HYBRIDDEPTH outperforms DFV by 5.9% in RMSE and 7.1% in AbsRel. Furthermore, compared to Defocus-Net [15], HYBRIDDEPTH achieves a 74.0% improvement in RMSE. These results highlight the effectiveness of fusing Depth Anything and DFV with our novel refinement layer.

Table 3 compares HYBRIDDEPTH with single-image depth estimation methods on the NYU Depth V2 dataset. The significant performance gaps above single-image approaches highlight the effectiveness of HYBRIDDEPTH. Specifically, compared to Depth Anything [29], HYBRIDDEPTH improves RMSE by 37.9%. It also outperforms diffusion-based models like ECoDepth [18] by 41.3% in RMSE. These results demonstrate that incorporating focal stack cues into the depth estimation task leads to substantial gains over single-image depth methods.

5.4. Zero-Shot Evaluation

We evaluated HYBRIDDEPTH's zero-shot performance, which is essential for most depth model applications, by comparing it against baselines on the ARKitScenes and Mobile Depth datasets. We show that HYBRIDDEPTH outperforms SOTA models on the ARKitScenes dataset and produces more detailed depth maps while preserving object boundaries on the Mobile Depth. Additionally, HYBRIDDEPTH demonstrates better consistency in depth es-

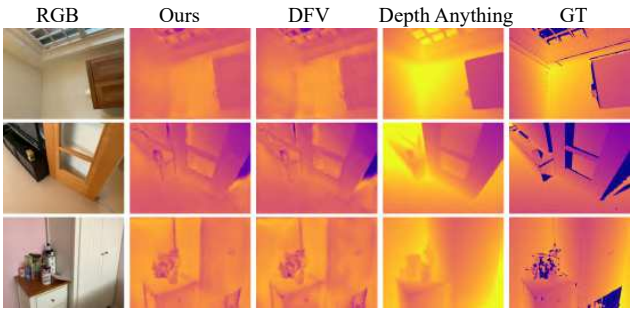

Figure 4. HYBRIDDEPTH's zero-shot performance on ARKitScenes compared to DFV and Depth Anything, demonstrating improved depth accuracy and detail preservation.

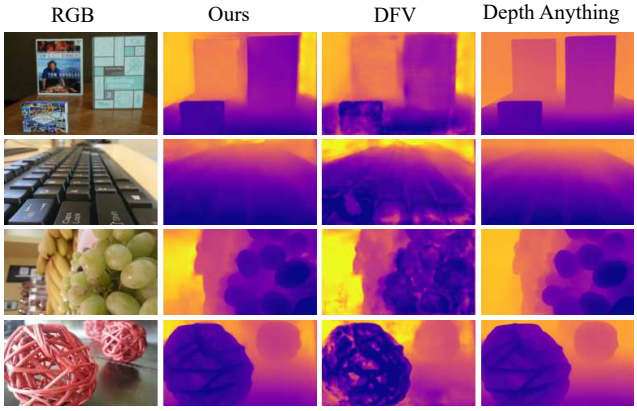

Figure 5. Qualitative results on Mobile Depth dataset.

timations across two different zoom levels, addressing the common scale ambiguity problem faced by single-image depth estimation models. Please check Supplementary Material for more visual comparisons of HYBRIDDEPTH to ARCore [1] and DFV.

Results on ARKitScenes. As shown in Table 4, HYBRIDDEPTH achieves a 32.6% improvement in RMSE over DFV and a 45.3% improvement over Depth Anything, while using a smaller model size. When trained on NYU Depth V2 dataset, HYBRIDDEPTH achieves the best RMSE results and comparable AbsRel performance, highlighting the effectiveness of focal stack cues in mobile AR scenarios.

Additionally, as seen in the comparison between the two versions of our model, one trained on NYU Depth V2 and the other on DDFF12, the model trained on NYU Depth V2 demonstrates better zero-shot performance, improving RMSE by 19.4% compared to the one trained on DDFF12. This suggests that training with more data, even synthetic ones like NYU Depth V2, can improve zero-shot performance in real-world scenarios. Figure 4 further illustrates HYBRIDDEPTH's performance in capturing details and producing high-quality depth maps. Also, note that Depth Anything overestimates the depth values for some regions.

Table 4. Zero-shot evaluation comparison on the ARKitScenes validation set with a focal stack size of 5. **Bold** represents the best results. Underline represents second best results. The unit for all metrics is meter.

Model	Type*	RMSE ↓	AbsRel ↓	#Params
ZoeDepth [4]†	SIDE	0.61	0.33	334.82M
DistDepth [27]	SIDE	0.94	0.45	68M
ZeroDepth [11]	SIDE	0.62	0.37	233M
Depth Anything [29]	SIDE	0.53	**0.32**	335.79M
DFV [28]	DFF	0.43	0.51	15M
Ours (NYU Depth V2)	DFF	**0.29**	0.42	65.6M
Ours (DDFF12)	DFF	0.36	0.49	65.6M

* *DFD* stands for depth from defocus/focus depth estimation. *SIDE* stands for single image depth estimation.
† For ZoeDepth we have used ZoeDepth-M12-N version.

Results on Mobile Depth. Figure 5 shows sample qualitative comparisons on Mobile Depth (other samples are shown in Supp.). We can see that HYBRIDDEPTH maintains higher levels of detail and object boundary accuracy. For example, in the last row, HYBRIDDEPTH captures the details on the ball correctly, but DFV's output is noisy. Comparing our results with Depth Anything, we can see that HYBRIDDEPTH successfully utilizes all the details captured by Depth Anything. More examples are available in Supplementary Material.

Effect of different zoom levels We evaluated HYBRIDDEPTH, DFV, and Depth Anything across two different zoom levels using two scenes (Figure 6). HYBRIDDEPTH maintains consistent depth estimations in both scenes. In the first scene (Figure 6a), HYBRIDDEPTH shows only a 5 cm difference between two zoom levels, while DFV exhibits a larger discrepancy of 18 cm and produces noisier results. Depth Anything significantly overestimates depth, with errors up to 240 cm, and shows inconsistencies of up to 120 cm between two zoom levels. In the second scene (Figure 6b), HYBRIDDEPTH achieves an impressive depth difference of just 3 cm, while Depth Anything has a 45 cm difference and DFV shows a 14 cm difference, both failing to provide a correct and consistent depth.

5.5. Ablation Study

We investigate the impact of several key design choices. See more ablation studies in Supplementary Material.

Effect of Uncertainty Map. Table 5 shows the performance comparison of HYBRIDDEPTH with and without the uncertainty map on the DDFF12 dataset. When the uncertainty map is included, model achieves an 8.6% reduction in MSE and a 2.4% reduction in RMSE. These improvements suggest that incorporating the uncertainty map helps the refinement layer to refine depth estimates more accurately, leading to more accurate estimations.

Effect of Refinement Layer. Table 6 shows the performance of HYBRIDDEPTH on the DDFF12 and NYU Depth

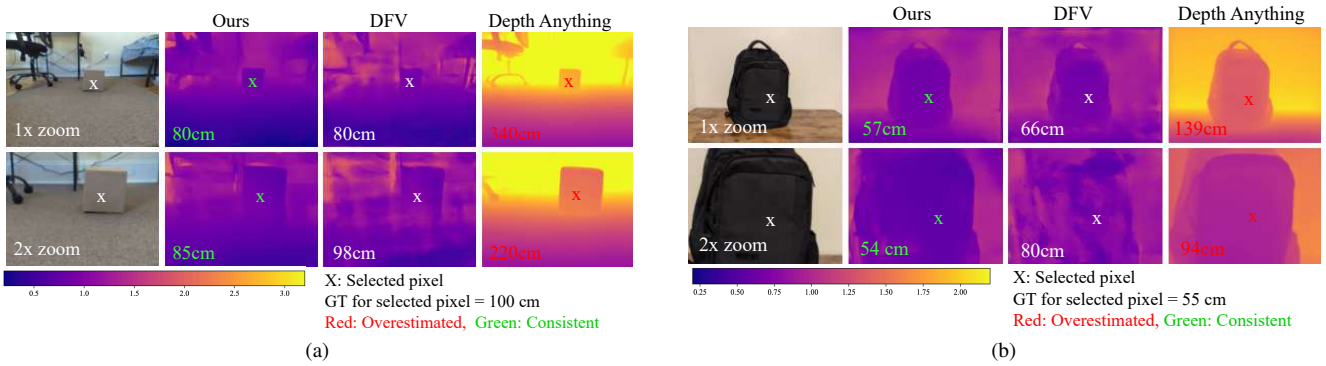

Figure 6. Comparison of depth estimations across different zoom levels for various models. HYBRIDDEPTH achieves more accurate and consistent estimations than DFV and Depth Anything. Specifically, DFV's estimations are more noisy and have larger discrepancies (e.g., 18 cm) between two zoom levels than HYBRIDDEPTH (e.g., 5 cm). Depth Anything vastly overestimates the depth of both scenes. We captured the images using our own focal stack capturing mobile app (§4.3) using a Google Pixel 6 Pro and obtained the GT depth using a measuring tape.

Table 5. Performance comparison between HYBRIDDEPTH trained with and without uncertainty map for refinement layer on DDFF12. The unit for all metrics is disparity.

Uncertainty Map	MSE ↓	RMSE ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
✗	5.6×10^{-4}	0.0205	0.168	0.78	0.94	0.98
✓	5.1×10^{-4}	0.0200	0.170	0.79	0.95	0.98

Table 6. Effect of refinement layer on NYU Depth V2 and DDFF12 with focal stack size of 5. The unit for DDFF12 is disparity and for NYU Depth V2 is meter.

Method	Dataset	RMSE ↓	AbsRel ↓	δ_1 ↑
Globally Scaled	DDFF12	0.0224	0.19	0.72
Globally Scaled + Refinement		0.0200	0.17	0.79
Globally Scaled	NYU Depth V2	0.552	0.18	0.77
Globally Scaled + Refinement		0.128	0.03	0.99

V2 datasets with and without the refinement layer, using a focal stack size of 5. The addition of the refinement layer improves depth estimation performance on both datasets. For DDFF12, the RMSE improves by 10.7% and AbsRel by 10.5%. For the NYU Depth V2 dataset, the refinement layer leads to a significant RMSE reduction of 76.8%, with AbsRel improving by 85.6%, demonstrating a substantial boost in accuracy.

6. Conclusion and Future Work

Achieving robust and accurate metric depth estimation in the wild remains a challenging and important problem. Recent work demonstrated that single-image depth models such as ZoeDepth have poor generalizability in challenging real-world environments [9]. Concurrently, we observe that existing foundational models like Depth Anything still face significant issues with scale ambiguity. Focal stack, data that has become increasingly available on mobile devices, has the potential to provide valuable depth information to address the scale ambiguity problem. These observations motivate our design of HYBRIDDEPTH, a novel end-to-end metric depth estimation pipeline.

At the core, HYBRIDDEPTH synergistically fuses metric information from the focal stack and depth prior from a foundational model, and uses a refinement model to further enhance details. HYBRIDDEPTH establishes new SOTA results on the commonly used DFF dataset DDFF12, improving RMSE and MSE over DFV [28] by 6.1% and 10.5%, respectively. On another real focal stack dataset, Mobile Depth, HYBRIDDEPTH achieves strong zero-shot performance. Similarly, on datasets with synthetic focal stack (NYU Depth V2 and ARKitScenes), HYBRIDDEPTH outperforms both DFF and single-image SOTA models. This robust performance is achieved with only mobile cameras, making HYBRIDDEPTH highly practical and accessible compared to solutions that rely on specialized hardware like LiDAR or ToF sensors.

As part of future work, we will explore improving both the DFF modules and the single-image depth prior. While we showcase stronger generalization capabilities compared to previous methods, there is still a substantial performance drop on out-of-domain samples. We would like to close the gap by scaling up training and better utilizing synthetic data.

Acknowledgement

This work was supported in part by NSF Grants #2350189, #2346133, and #2236987.

References

[1] Build new augmented reality experiences that seamlessly blend the digital and physical worlds. https://developers.google.com/ar. Accessed: 2024-5-8. 7, 11

[2] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad 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, 11

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *ECCV*, 2022. 2

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv:2302.12288*. 1, 2, 3, 6, 7, 11

[5] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1–a model zoo for robust monocular relative depth estimation. *arXiv:2307.14460*, 2023. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 11

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. 4

[8] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth Estimation Using Adaptive Bins. In *CVPR*, 2021. 2

[9] Ashkan Ganj, Yiqin Zhao, Hang Su, and Tian Guo. Mobile AR Depth Estimation: Challenges & Prospects. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, HOTMOBILE '24, page 21–26, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2, 5, 8

[10] Shiva Ghasemi, Majid Behravan, Sunday D. Ubur, and Denis Gračanin. Attention and sensory processing in augmented reality: Empowering adhd population. In Margherita Antona and Constantine Stephanidis, editors, *Universal Access in Human-Computer Interaction*, pages 301–320, Cham, 2024. Springer Nature Switzerland. 1

[11] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rares Ambrus, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *ICCV*, 2023. 7

[12] 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. 2, 6, 11

[13] David Jones, Shiva Ghasemi, Denis Gračanin, and Mohamed Azab. Privacy, safety, and security in extended reality: User experience challenges for neurodiverse users. In Abbas Moallem, editor, *HCI for Cybersecurity, Privacy and Trust*, pages 511–528, Cham, 2023. Springer Nature Switzerland. 1

[14] Nazanin Mahjourian and Vinh Nguyen. Multimodal object detection using depth and image data for manufacturing parts, 2024. 1

[15] Maxim Maximov, Kevin Galim, and Laura Leal-Taixe. Focus on defocus: Bridging the synthetic to real domain gap for depth estimation. In *CVPR*, 2020. 2, 5, 6

[16] Michael Moeller, Martin Benning, Carola Schönlieb, and Daniel Cremers. Variational depth from focus reconstruction. *IEEE Transactions on Image Processing*, 24(12):5369–5378, 2015. 6

[17] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 11

[18] Suraj Patni, Aradhye Agarwal, and Chetan Arora. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. *arXiv preprint arXiv:2403.18807*, 2024. 1, 6

[19] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 2

[20] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 2, 3

[21] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 4

[22] Haozhe Si, Bin Zhao, Dong Wang, Yupeng Gao, Mulin Chen, Zhigang Wang, and Xuelong Li. Fully self-supervised depth estimation from defocus clue. *arXiv preprint arXiv:2303.10752*, 2023. 4, 5

[23] Supasorn Suwajanakorn, Carlos Hernandez, and Steven M. Seitz. Depth from focus with your mobile phone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 11, 12

[24] Ning-Hsu Wang, Ren Wang, Yu-Lun Liu, Yu-Hao Huang, Yu-Lin Chang, Chia-Ping Chen, and Kevin Jou. Bridging unsupervised and supervised depth from focus via all-in-focus supervision. In *ICCV*, 2021. 2

[25] 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, 4

[26] Changyeon Won and Hae-Gon Jeon. Learning depth from focus in the wild, 2022. 6

[27] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. Toward practical monocular indoor depth estimation. In *CVPR*, 2022. 7

[28] Fengting Yang, Xiaolei Huang, and Zihan Zhou. 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, pages 12632–12641, United States, 2022. IEEE Computer Society. Publisher Copyright: © 2022 IEEE.; 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022 ; Conference date: 19-06-2022 Through 24-06-2022. 1, 2, 3, 6, 7, 8, 11

[29] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2, 3, 6, 7, 11

[30] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. 2023. 2

[31] Xiaze Zhang, Ziheng Ding, Qi Jing, Yuejie Zhang, Wenchao Ding, and Rui Feng. Deeppointmap: Advancing lidar slam with unified neural descriptors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10413–10421, 2024. 1

[32] Yunfan Zhang, Tim Scargill, Ashutosh Vaishnav, Gopika Premsankar, Mario Di Francesco, and Maria Gorlatova. Indepth: Real-time depth inpainting for mobile augmented reality. *IMWUT*, 2022. 3

[33] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *ICCV*, 2023. 6

A. Implementation

A.1. Training Details

For the NYU Depth V2 dataset, we set weight decay (λ) to 0.001 and use a learning rate of 3×10^{-4}. The batch size is 24, and we use the original data size (480×640) without any resizing.

For DDFF12, weight decay (λ) is set to 0.0001, with a learning rate of 1×10^{-4}. The batch size is 8, and the input size during training is 224×224 pixels, with random crop and flip augmentations applied. For evaluation, the original image size of 383×552 is used, following DFF-based methods [28]. Focal stacks are arranged in ascending order of focal distance to ensure consistency in depth processing.

For the refinement layer, we initialize the MiDaS-small encoder backbone with pre-trained ImageNet [6] weights, while the remaining layers are randomly initialized to allow adaptation to our depth estimation task.

B. Experiments

B.1. Dataset

DDFF12 [12]. We follow the dataset split specified in DFV [28]. The training set consists of six scenes, each containing 100 samples, while the test set includes six different scenes with 20 samples per scene. Each sample contains a 10-frame focal stack along with a corresponding ground truth disparity map. The images have a resolution of 383 × 552 pixels. For our training and evaluation, we used a focal stack of 5 frames, similar to DFV [28].

Mobile Depth [23] includes 11 aligned focal stacks from 11 different scenes. The image resolutions range from 360 × 640 to 518 × 774, with each stack containing between 14 and 33 frames. Since ground truth depth and focal distance are not provided, we used this dataset solely for qualitative comparisons on aligned focal stack images.

NYU Depth V2 [17] contains over 24K densely labeled RGB and depth image pairs in the training set and 654 pairs in the test set. This dataset covers a broad range of indoor environments, with ground truth depth maps obtained using a structured light sensor, provided at a resolution of 640 × 480 pixels.

ARKitScenes [2] is a large-scale dataset designed for mobile AR applications. For our experiments, we utilized a subset of 5.6K images for evaluating HYBRIDDEPTH's zero-shot performance. This subset provides a comprehensive basis for evaluating the robustness and accuracy of our model under real-world AR conditions.

B.2. Model Performance Analysis

We conducted a performance analysis to demonstrate the efficiency of our model compared to SOTA models like ZoeDepth-M12-N, Depth Anything, and DFV. All tests were performed on an Nvidia RTX 4090 GPU. Table 7 shows that HYBRIDDEPTH achieves an inference time of 20 ms, which is 4.3X faster than ZoeDepth-M12-N and 2.85X faster than Depth Anything. Additionally, our model's size is 5.3X smaller than ZoeDepth-M12-N and 5.2X smaller than Depth Anything. Despite being more compact, HYBRIDDEPTH provides a considerable improvement in performance and is highly suitable for deployment on devices with limited memory and storage. While DFV is faster at 8 ms and smaller in size, previous sections have shown that its depth estimation accuracy is significantly lower.

Table 7. Performance analysis of the three SOTA models on Nvidia RTX 4090 with DDFF12. Note: We use the ViT Large version for Depth Anything.

Model	Inference Time	Size	#Params
ZoeDepth-M12-N [4]	86 ± 6 ms	1.28 GB	344.82M
Depth Anything [29]	57 ± 5 ms	1.25 GB	335.79M
DFV	8 ± 2 ms	0.07 GB	15M
Ours	20 ± 2 ms	0.24 GB	65.6M

B.3. Qualitative Comparison

Qualitative Comparison with ARCore and DFV. Depth estimation plays a crucial role in augmented reality (AR) applications, where accurate depth maps are essential for tasks such as rendering occlusions and precise object placement. We compared our model against the depth maps generated by the commercial ARCore framework [1] and DFV [28]. Utilizing an Android app, we captured a focal stack of five images and sent it over WiFi to an edge server for alignment and inference. Figure 7 shows that our model preserves better edge details and object boundaries compared to ARCore, while also producing smoother and more reliable depth maps than DFV.

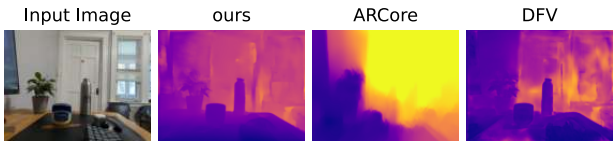

Figure 7. Qualitative comparison with ARCore and DFV. Our model outputs better depth by preserving object boundaries and overall geometrical information about the scene. In our experiments with ARCore, depth maps were obtained by moving the camera around the scene until no further improvement was observed.

Qualitative Comparison on Mobile Depth. Figure 8 presents additional results on the aligned scenes of the Mobile Depth dataset. All deep learning methods generalize well to these scenes without fine-tuning. In row 4, our method successfully captures intricate details in the plants,

and in the last row, HYBRIDDEPTH provides smoother and more accurate depth estimations, even capturing the depth behind objects. However, our model struggles with depth estimation for transparent surfaces, such as glass. The focal stacks in rows 6,7 are taken from the same scenes with different camera motions, therefore have slightly different frame alignment. We refer readers to [23] for more details of this dataset. Overall, HYBRIDDEPTH consistently delivers smoother depth maps with better boundary preservation compared to other methods.

Qualitative Comparison on NYU Depth V2. Figure 9 compares our model with Depth Anything on the NYU Depth V2 dataset. Both models generate accurate depth maps; however, our model excels at capturing depth for distant objects more closely aligned with the ground truth, as seen in rows 3 and 6. Additionally, our model captures finer details more effectively, particularly in row 2.

B.4. Ablation Study

Effect of Focal Stack Size. We analyzed the effect of focal stack size on HYBRIDDEPTH's performance across the NYU Depth V2, DDFF12, and ARKitScenes datasets (Table 8). On the NYU Depth V2 dataset, increasing the focal stack size from 5 to 10 reduced the RMSE by 35.2% and the AbsRel by 42.3%, while both configurations still achieved state-of-the-art (SOTA) results. Similarly, on the ARKitScenes dataset, using a focal stack size of 10 slightly reduced the RMSE by 10.3%, confirming that HYBRIDDEPTH's performance benefits from a larger focal stack size but remains robust even with smaller stacks. The performance difference on the DDFF12 dataset was negligible between stack sizes, demonstrating consistent accuracy across different configurations.

Table 8. Effect of focal stack size on HYBRIDDEPTH. Both focal stack sizes yield new SOTA results, and there are no significant performance differences between these two settings.

Focal Stack Size	Trained	Evaluated	RMSE ↓	AbsRel ↓
5	NYU Depth V2	NYU Depth V2	0.128	0.026
10	NYU Depth V2	NYU Depth V2	0.083	0.015
5	DDFF12	DDFF12	0.0200	0.1695
10	DDFF12	DDFF12	0.0200	0.1690
5	NYU Depth V2	ARKitScenes	0.29	0.42
10	NYU Depth V2	ARKitScenes	0.29	0.39

Different Global Scaling Methods. We evaluated the performance of various global scaling (GS) methods on the DDFF12 dataset, as shown in Table 9. The least square method showed competitive performance, achieving results comparable to more complex method RANSAC, but with a significant computational advantage. For example, it was over 30x faster than RANSAC with 200 iterations and 50 sample size, while providing similar accuracy with only a 1.8% increase in RMSE compared to the best RANSAC configuration. This makes the least square method the most efficient choice for global scaling, ensuring reliable depth estimates without adding considerable overhead.

Table 9. Comparison of global scaling (GS) methods on the DDFF12 dataset.

Method	RMSE ↓	AbsRel ↓	δ_1 ↑	Time (ms) ↓
Least Square	0.0224	0.19	0.72	3
RANSAC (itr: 60, Sample size: 5)	0.0246	0.19	0.73	34
RANSAC (itr: 100, Sample size: 20)	0.0236	0.18	0.76	96
RANSAC (itr: 200, Sample size: 50)	0.0228	0.17	0.75	170

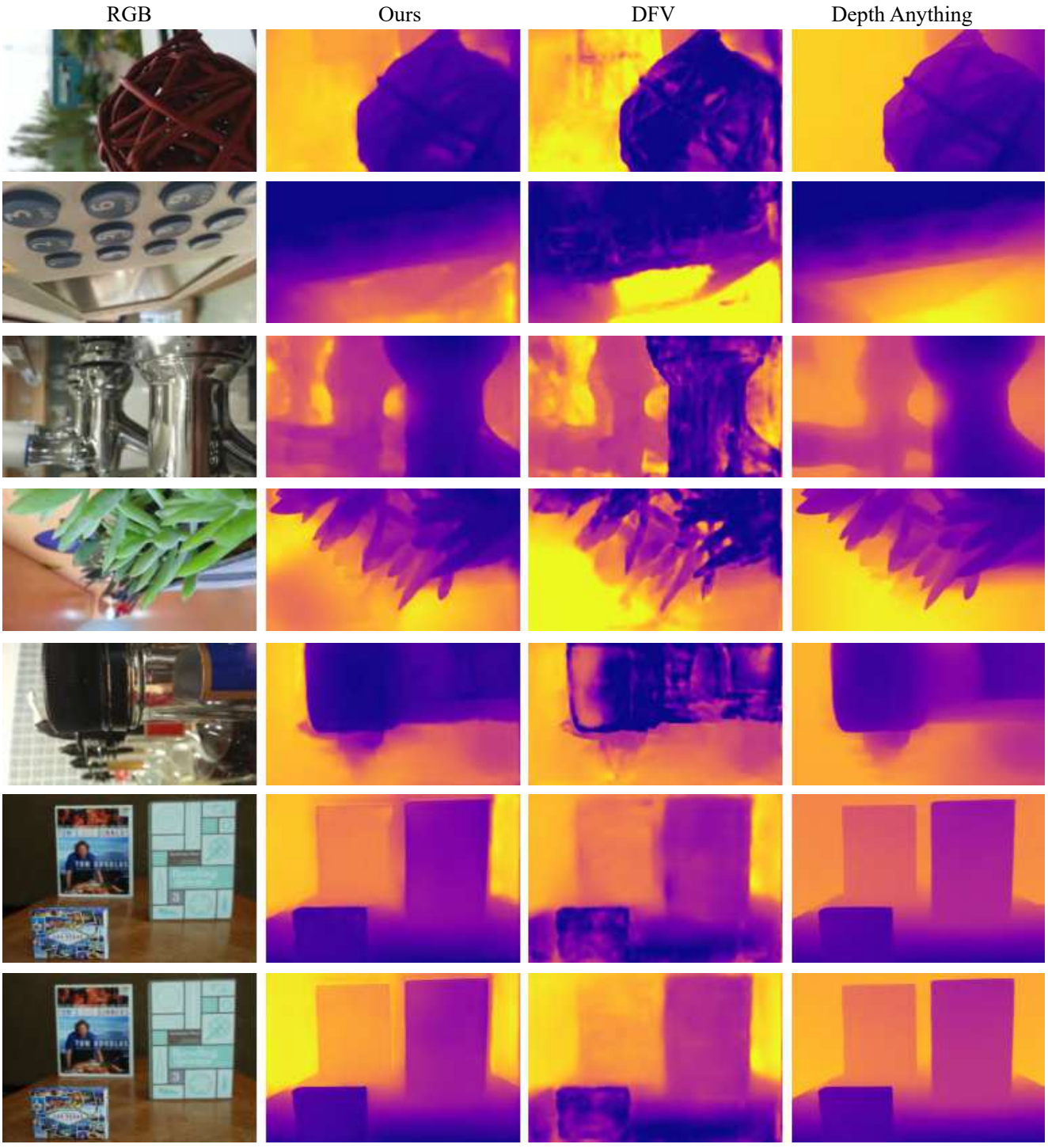

Figure 8. Additional qualitative results on the Mobile Depth dataset. The focal stacks in rows 6,7 are taken from the same scenes with different camera motions, therefore have slightly different frame alignment

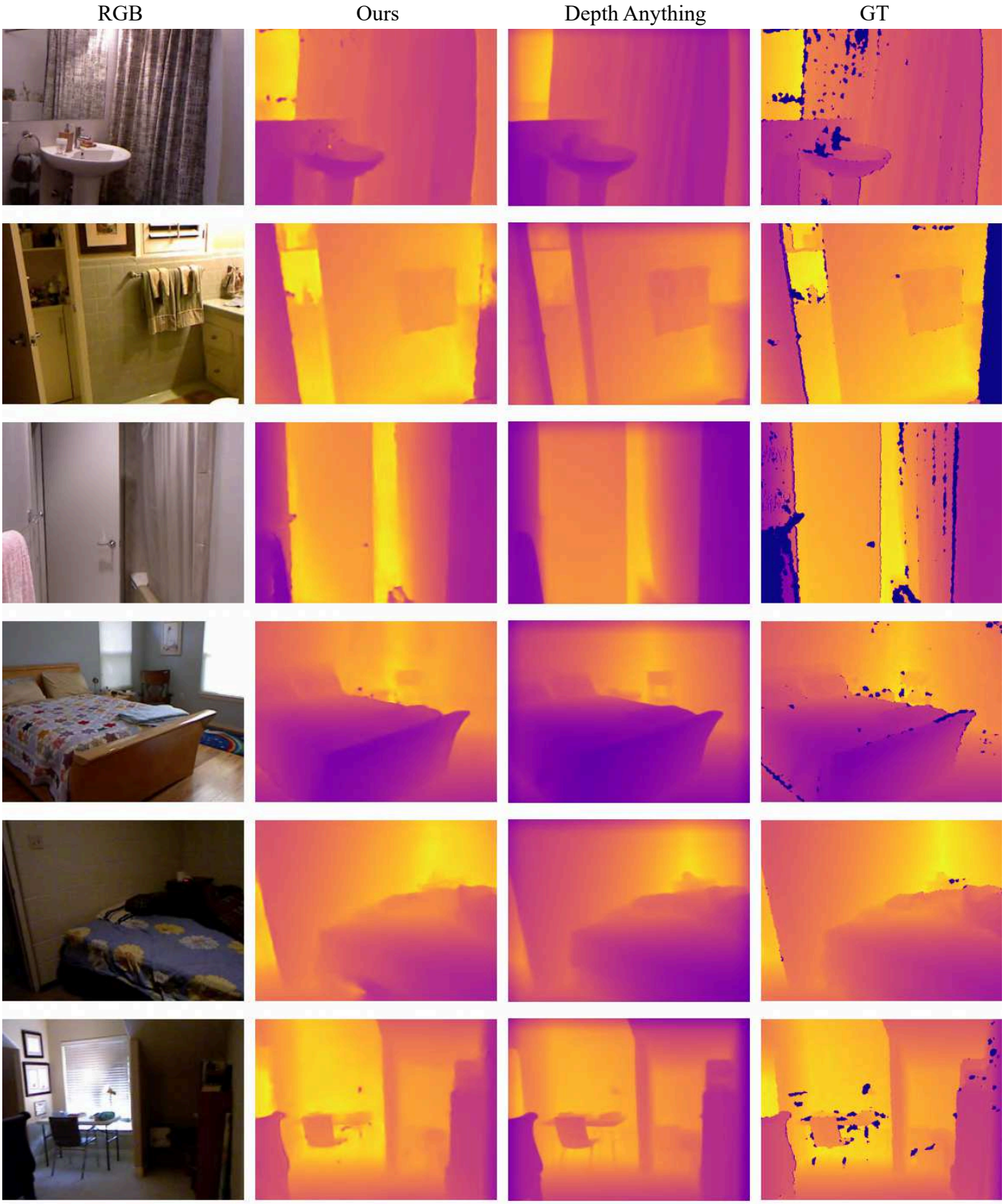

Figure 9. Qualitative results on the NYU Depth V2 dataset.