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 [27]. 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 [27]. 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 [27].

Mobile Depth [22] 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 [16] 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 [28]	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 [27]. 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.

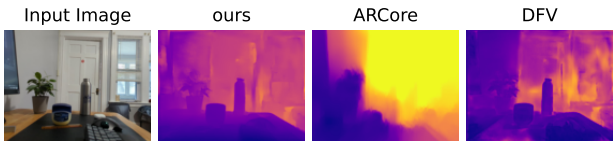

Figure 7. Qualitative comparison with ARCore and DFV. Our model outputs better depth by preserving object boundaries and overall geometrical information about the scene. In our experiments with ARCore, depth maps were obtained by moving the camera around the scene until no further improvement was observed.

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 [22] 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

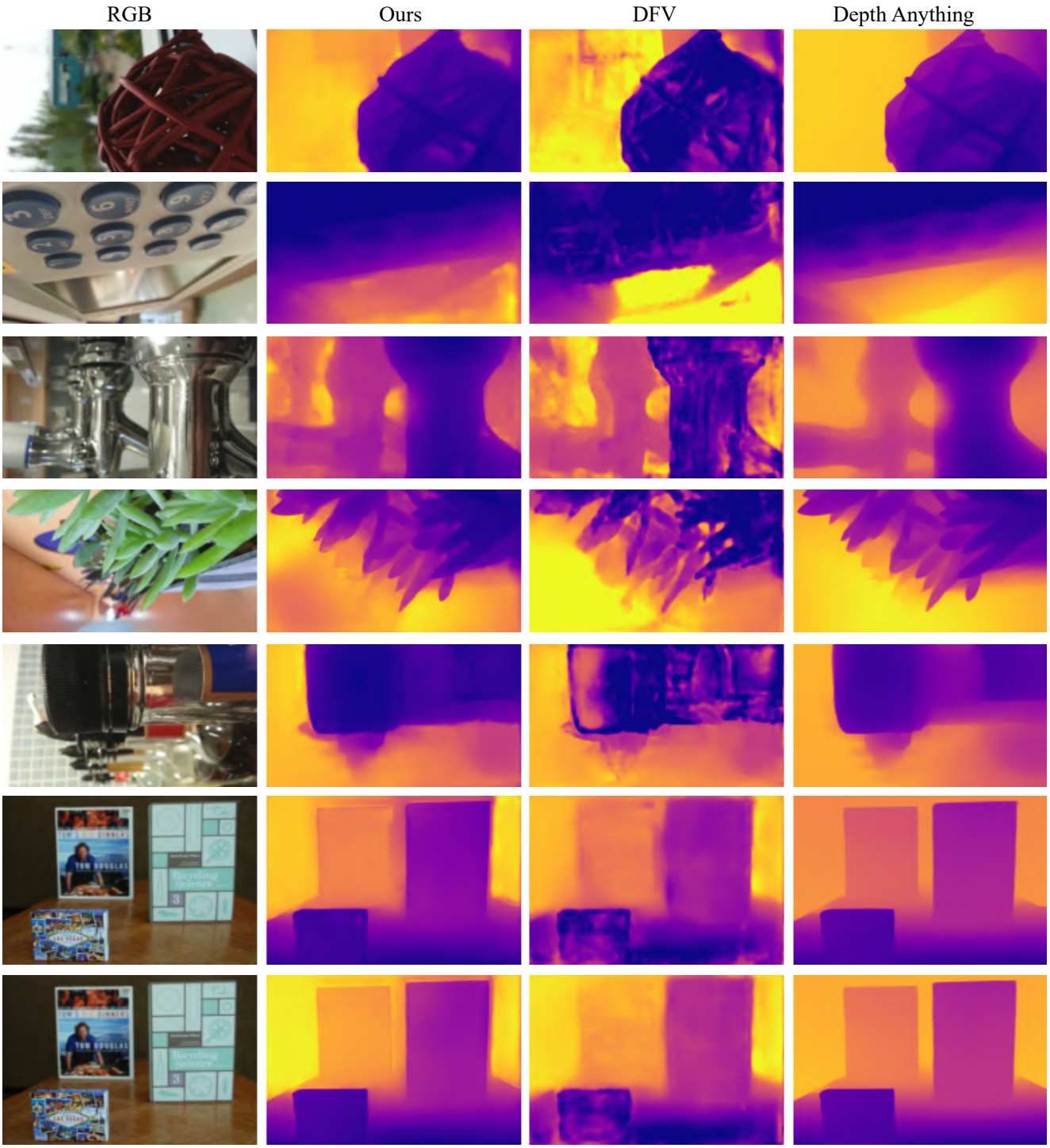

Figure 8. Additional qualitative results on the Mobile Depth dataset. The focal stacks in rows 6,7 are taken from the same scenes with different camera motions, therefore have slightly different frame alignment

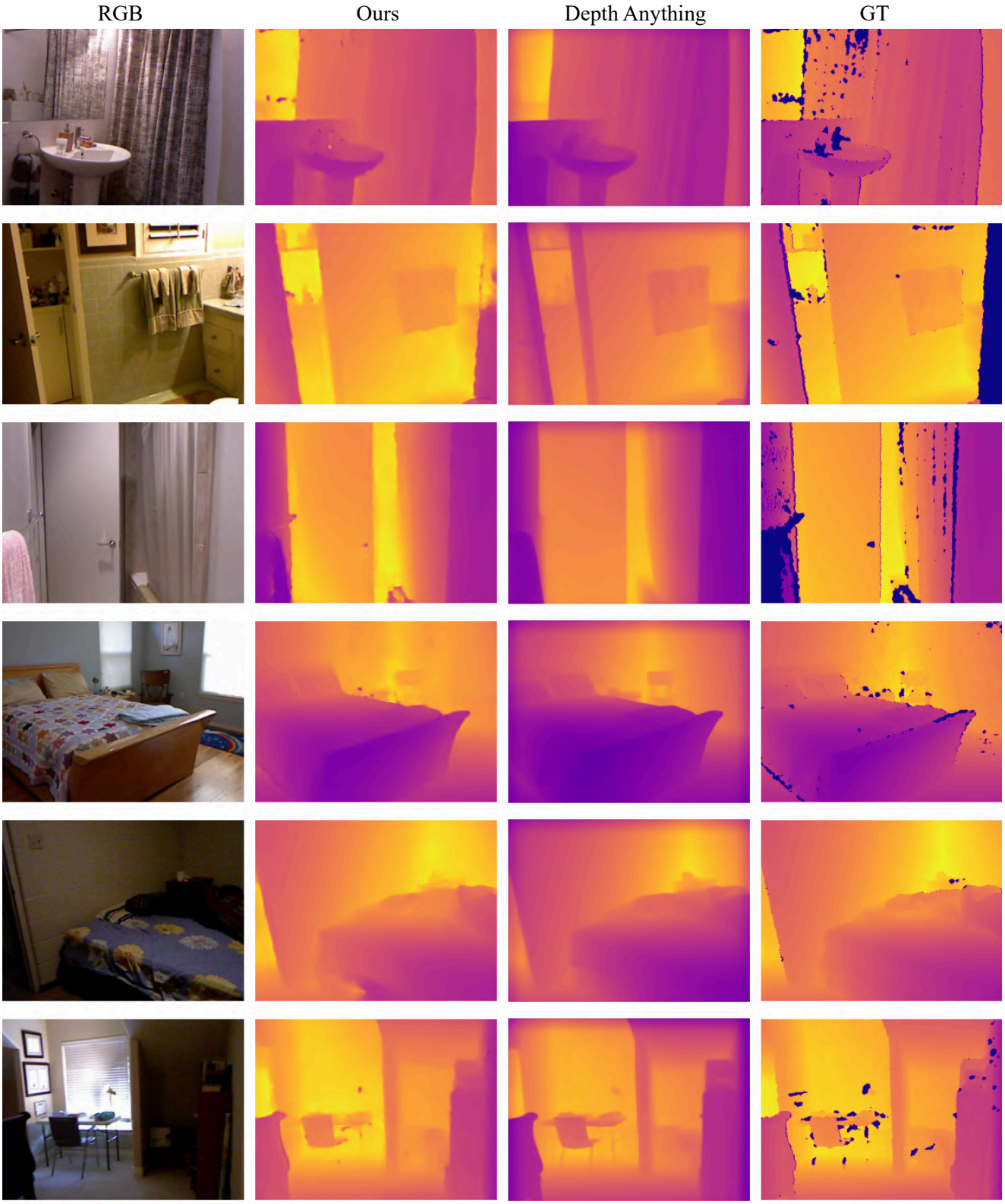

Figure 9. Qualitative results on the NYU Depth V2 dataset.