

GeoDiff: Geometry-Guided Diffusion for Metric Depth Estimation

Tuan Pham^{*} Thanh-Tung Le^{*} Xiaohui Xie[†] Stephan Mandt[†]
University of California, Irvine ^{*} Equal contribution [†] Equal advising
{tuan.pham, tung.le, xhx, mandt}@uci.edu

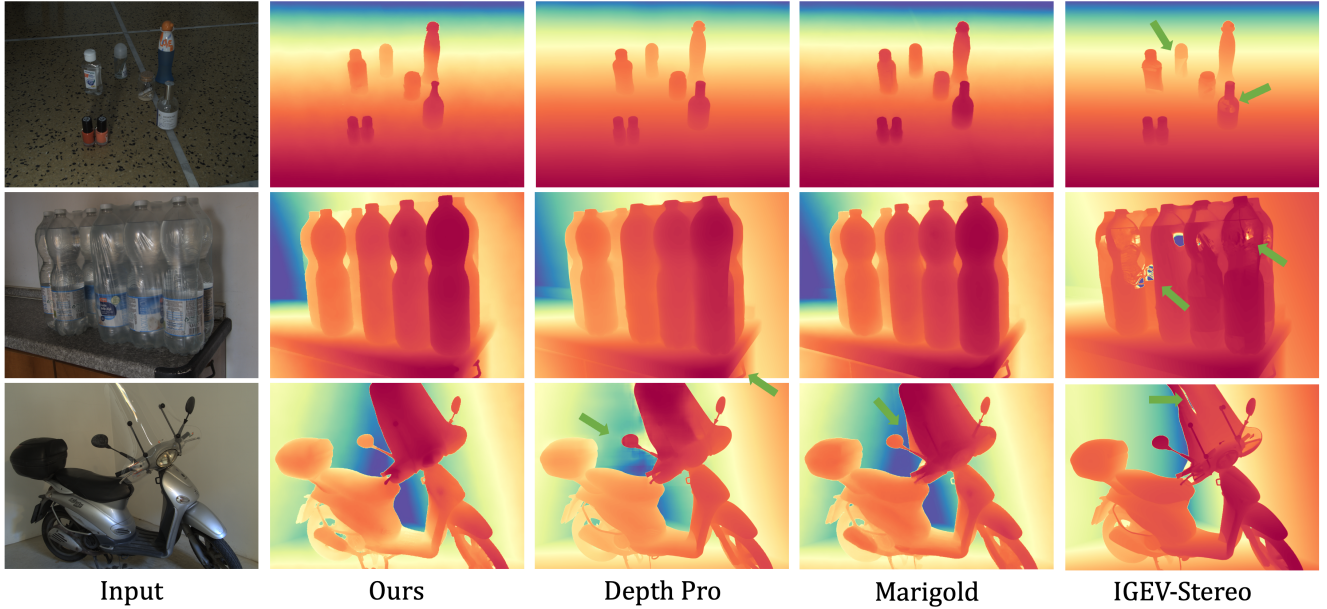


Figure 1. **Results on the challenging Booster [36] dataset.** Comparison of depth estimation performance on transparent and reflective objects. From left to right: input image, followed by depth maps from GeoDiff (ours), DepthPro [5], Marigold [23], and IGEV-Stereo [58]. Our method generates sharp, accurate metric depth maps in a zero-shot setting, leveraging stereo pairs for improved depth recovery.

Abstract

specular surfaces, all without requiring retraining.

We introduce a novel framework for metric depth estimation that enhances pretrained diffusion-based monocular depth estimation (DB-MDE) models with stereo vision guidance. While existing DB-MDE methods excel at predicting relative depth, estimating absolute metric depth remains challenging due to scale ambiguities in single-image scenarios. To address this, we reframe depth estimation as an inverse problem, leveraging pretrained latent diffusion models (LDMs) conditioned on RGB images, combined with stereo-based geometric constraints, to learn scale and shift for accurate depth recovery. Our training-free solution seamlessly integrates into existing DB-MDE frameworks and generalizes across indoor, outdoor, and complex environments. Extensive experiments demonstrate that our approach matches or surpasses state-of-the-art methods, particularly in challenging scenarios involving translucent and

1. Introduction

Depth estimation is an essential task and play a fundamental role in wide applications, such as 3D reconstruction [25, 33], autonomous driving [55], and AI-generated content [29, 64]. Recently, monocular depth estimation (MDE) [4, 12, 13, 42] and stereo depth estimation (StDE) have emerged as the leading methods for depth estimation. MDE approaches generally predict relative depth, which is invariant to scale and shift, whereas StDE methods focus on predicting disparity between two input images, which can be converted to metric depth (in meters) using known camera baseline and focal length. While recent methods attempt direct metric depth prediction from monocular images through large foundation models [60, 61, 63], these approaches demand extensive synthetic and real data and

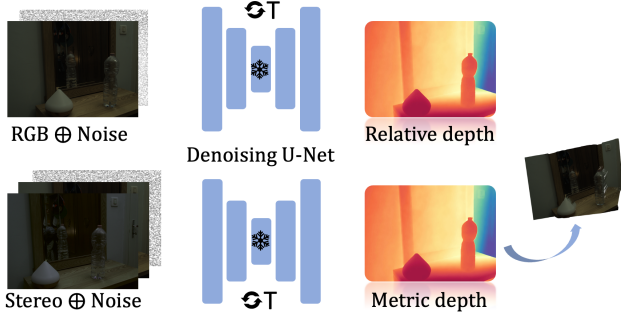


Figure 2. **Top.** Prior methods [14, 23] focus on fine-tuning diffusion models to estimate relative depth. **Bottom.** In contrast, our approach, without retraining, combines a pretrained monocular model with geometric guidance from stereo cues to directly predict metric depth in meters, achieving superior accuracy in even challenging scenes such as transparent and reflective surfaces.

are computationally expensive to train.

Recently, diffusion models have demonstrated potential as robust priors for zero-shot dense prediction tasks, including depth estimation [11, 14, 21, 23, 34, 42, 43]. Marigold [23] is among one of the pioneer methods that propose to repurpose diffusion-based image generators for MDE. The main idea is to finetune pretrained latent diffusion models (LDMs) [41], which have been trained on extensive text-to-image datasets, to generate depth maps from noise by conditioning on RGB images. The simple diffusion-based MDE (DB-MDE) paradigm works surprisingly well, delivering strong performance across a diverse range of natural images. Subsequent studies [11, 14, 17, 20] have developed upon this paradigm, establishing diffusion-based MDE as an active research in dense prediction task. However, estimating metric depth from a single image remains an inherently ill-posed and challenging problem, leading most DB methods to concentrate on reconstructing relative depth rather than absolute metric depth.

In this work, we advance this line of research by leveraging pretrained priors from DB-MDE models to achieve metric depth estimation through the incorporation of additional stereo settings. Specifically, we reformulate the depth sampling process as an inverse problem (IP) solved through diffusion models [7, 8, 10, 22, 45]. Our approach leverages pretrained LDMs, along with stereo vision-based geometric guidance, to learn the scale and shift for any given scene. Built upon the foundation of diffusion-based MDE (DB-MDE) approaches, such as Marigold [23], our method is scene-agnostic—applicable to objects, indoor, and outdoor scenes—and can be integrated with any DB-MDE framework following a similar schema. Extensive experiments across diverse datasets demonstrate that our approach performs effectively on a wide range of scenes and depth scales without requiring re-training for specific use cases. In summary, we propose:

- A novel framework that leverages diffusion-based MDE priors and stereo settings to achieve metric depth estimation.
- An IP-based approach for depth estimation that introduces a plug-and-play module, seamlessly integrating with any pretrained diffusion-based depth models that use iterative updates.
- Extensive experimental evaluations of our method compared to other competing methods across various datasets, encompassing indoor, outdoor, and challenging scenes with translucent or specular surfaces.

2. Related works

2.1. Depth Estimation

Depth estimation has various applications in 3D vision [14, 18, 19, 23, 26, 27, 50]. Monocular depth estimation and stereo matching have both seen significant advancements in recent years. Traditional monocular methods focused on in-domain metric depth estimation but faced challenges in generalization, leading to a shift towards zero-shot relative depth estimation using approaches like Stable Diffusion [41] for depth denoising and large-scale datasets such as [14, 23]. However, these data-driven models are limited by the availability of synthetic dataset, which has prompted methods like DepthAnything [60, 61], DepthPro [5] to leverage vast amounts of additional real dataset for enhanced robustness. Stereo matching, on the other hand, relies heavily on cost volume filtering techniques, typically employing deep learning models to extract features, build cost volumes, and regress disparities. Models like GCNet [24] and PSMNet [6] use 3D CNN architectures to address challenges with occlusions and textureless surfaces, while newer methods, such as GwcNet [18] and ACVNet [57], introduce group-wise correlation and attention mechanisms to improve cost volume expressiveness. RAFT-Stereo [30] adapts the optical flow network RAFT [51] with multi-level convolutional GRUs, achieving impressive results. Building on this foundation, IGEV-Stereo [58] proposes iterative geometry encoding volumes, demonstrating improved robustness. Despite these advancements, the high memory and computational costs of 3D convolutions limit scalability. Our method leverages diffusion priors and geometric guidance, effectively overcoming the limitations of traditional cost volume approaches, particularly for challenging surfaces like transparent or reflective regions. In particular, our approach employs pretrained model from diffusion-based MDE combining with stereo information to estimate metric depth in the wild. To the best of our knowledge, we are the first to employ a training-free approach that combines both monocular and stereo information to solve metric depth estimation effectively.

2.2. Diffusion Models for Inverse Problem Solving

Inverse problems (IPs) are ubiquitous and associated with a wide range of reconstruction problems such as computational image [1, 2], medical imaging [39, 49], and remote sensing [31]. IPs aim to recover an unknown sample $x \in \mathbb{R}^n$, given observed measurements $y \in \mathbb{R}^m$ of the form: $y = \mathcal{A}(x) + e$, where function $\mathcal{A}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the forward measurement operator and $e \in \mathbb{R}^m$ is additive noise. In the literature, the traditional approach of using hand-crafted priors (e.g. sparsity) is slowly being replaced by rich, learned priors such as diffusion generative models. While recent works [7, 8, 48] propose to solve inverse problem in pixel space, which is computationally expensive, authors [8, 45] recently introduce a method to solve IP in the latent space. In this work, we demonstrate that by using latent DM and classical stereo vision as geometry guidance, we can solve for metric depth estimation problem without re-training MDE. We hope that our approach potentially opens a new direction for tackling depth estimation problems.

3. Background

In this section, we provide background regarding diffusion models for monocular depth estimation in Section 3.1. Then, we describe how to employ diffusion models for solving inverse problems in Section 3.2. Finally, we demonstrate the differentiable warping module in Section 3.3.

3.1. Diffusion models for MDE

MDE is formulated as a conditional denoising diffusion generation task, modeling $p(x|y)$, where $x \in \mathbb{R}^{H \times W \times 1}$ denotes depth and $y \in \mathbb{R}^{H \times W \times 3}$ represents RGB input. The forward process progressively perturbs data via Gaussian kernels through a variance-preserving SDE [47]:

$$dx = -\frac{\beta_t}{2}xdt + \sqrt{\beta_t}dw \quad (1)$$

where $\beta_t \in (0, 1)$ is the monotonically increasing noise schedule and w denotes standard Wiener process. The reverse process learns the corresponding reverse SDE:

$$dx = \left[-\frac{\beta_t}{2}x - \beta_t \nabla_{x_t} \log p(x_t|y) \right] dt + \sqrt{\beta_t}d\bar{w} \quad (2)$$

where $\nabla_{x_t} \log p(x_t)$ is the score function and $d\bar{w}$ denotes backward Wiener process. A denoising score matching network [53] is trained to approximate the score function:

$$\hat{\theta} = \operatorname{argmin}_{\theta} \mathbb{E}[\|s_{\theta}(x_t, y, t) - \nabla_{x_t} \log p(x_t|x_0, y)\|_2^2] \quad (3)$$

The trained score function s_{θ} is then used to approximate the reverse-time SDE through numerical simulation.

3.2. Diffusion Models for Solving Inverse Problems

In inverse problems, we can recover an unknown signal x from measurements y related by $y = \mathcal{A}(x) + e$, where $\mathcal{A}(\cdot)$ is the forward measurement operator and $e \sim \mathcal{N}(0, \sigma^2 I)$ represents Gaussian noise [7]. Applying Bayes' theorem to the conditional score:

$$\nabla_{x_t} \log p(x_t | y) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(y | x_t) \quad (4)$$

Under mild assumptions [7], we approximate:

$$\nabla_{x_t} \log p(y | x_t) \approx \nabla_{x_t} \log p(y | \hat{x}_0) \quad (5)$$

where $\hat{x}_0$ is the one-step prediction via Tweedie's formula [40]:

$$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} (x_t + \sqrt{1 - \bar{\alpha}_t} s_{\theta}(x_t, t, y_1)) \quad (6)$$

With Gaussian noise, we derive:

$$\nabla_{x_t} \log p(y | \hat{x}_0) = -\frac{1}{\sigma^2} \nabla_{x_t} \|y - \mathcal{A}(\hat{x}_0)\|_2^2 \quad (7)$$

The final conditional score becomes:

$$\nabla_{x_t} \log p(x_t | y) = \nabla_{x_t} \log p(x_t) - \lambda \nabla_{x_t} \|y - \mathcal{A}(\hat{x}_0)\|_2^2 \quad (8)$$

where λ controls the strength of additional guidance to the original score function.

3.3. Differentiable warping

Given a stereo pair (or two views with known poses) y_1, y_2 and the corresponding depth maps x_1, x_2 ; we define an operation that projects each pixel from the source image y_1 onto the target image y_2 . Using the intrinsic matrices $K_1, K_2 \in \mathbb{R}^{3 \times 3}$ of the source and target cameras, respectively, and the relative transformation $T_{1 \rightarrow 2} \in \mathbb{R}^{4 \times 4}$ between the cameras, the forward warping is formulated as:

$$c_2 \sim K_2 T_{1 \rightarrow 2} x_1(c_1) K_1^{-1} c_1 \quad (9)$$

where c_1 and c_2 denote the homogeneous pixel coordinates in y_1 and y_2 , respectively, and $x_1(c_1)$ represents the depth at pixel c_1 in y_1 .

Based on this coordinates mapping, we can define the forward warping operator $P_{y_1 \rightarrow y_2}(x_1, y_1)$, which projects y_1 onto y_2 ; and the backward warping operator $P_{y_2 \rightarrow y_1}(x_1, y_2)$. Notably, both warping operations rely solely on the depth map x_1 from y_1 . Backward warping, in particular, has been extensively applied for computing reprojection losses in self-supervised depth estimation [15, 16] or online stereo depth adaptation [52, 65]. We provide in-depth discussion in the Supplementary 7.2.

A detailed derivation can be found in our Supplemental 7.1.

However, computing the forward warping function $P_{y_1 \rightarrow y_2}$ is not preferred due to holes and computational complexity (see Supplemental 7.2). Therefore, we adopt a backward warping function $P_{y_2 \rightarrow y_1}$ instead of a forward warping. We also follow previous works [15, 16] to employ a linear combination of SSIM and L1 losses to calculate the geometric reprojection loss:

$$\mathcal{L}_{geo} = \eta(1 - \text{SSIM}(y_1, P_{y_2 \rightarrow y_1}(\tilde{x}_0, y_2)))/2 + (1 - \eta)\|y_1 - P_{y_2 \rightarrow y_1}(\tilde{x}_0, y_2)\|_1 \quad (14)$$

where η is the hyperparameter that balances the two losses. Our final score function is:

$$\nabla_{z_t} \log p(z_t | y_1, y_2) = s_\theta(z_t, t, y_1) - \lambda \nabla_{z_t} \mathcal{L}_{geo}(y_1, P_{y_2 \rightarrow y_1}(\tilde{x}_0, y_2)) \quad (15)$$

At every diffusion sampling step, we leverage this score function to update the current latent z_t , while simultaneously updating the scale $\hat{s}$ and shift $\hat{t}$ using the reprojection gradient. The detailed algorithm is shown in the Supplemental 8.

Generalization. Acquisition of stereo image pairs in unconstrained real-world scenarios presents significant practical challenges, typically necessitating a calibrated dual-camera setup with precise side-by-side alignment. Our proposed framework, however, extends beyond traditional stereo configurations to accommodate arbitrary two-view settings with known relative poses, substantially enhancing its applicability across diverse deployment contexts. Our method only requires relative transformation for performing differentiable warping operations (detailed in Supplemental 7.2). For in-the-wild image pairs lacking calibration metadata, we leverage recent advances in foundation models for dense 3D reconstruction [54] to estimate the requisite camera intrinsic and extrinsic parameters.

Regularization. To improve the stability of our optimization, we introduce a global scale hyperparameter g_s and regularization loss for scale $\hat{s}$ and shift $\hat{t}$. Through our experiments, we observe that each scene in the wild has different depth scale. The further the depth is, the better the warped image can be rendered, thus minimizing the $\mathcal{L}_{geo}$, and keep enforcing the depth further away. It leads to the wrong sampling trajectory of the diffusion model. Therefore, we opt to pre-select a global scale g_s and also apply L_2 regularize on the $\hat{s}$ and shift $\hat{t}$ parameters. Now, the $\tilde{x}_0$ and the total optimization loss $\mathcal{L}_{geo}$ become:

$$\tilde{x}_0 = g_s[\text{softplus}(\hat{s}) \cdot \hat{x}_0 + \text{softplus}(\hat{t})] \quad (16)$$

$$\mathcal{L}_{geo} = \mathcal{L}_{geo} + \gamma(\|\hat{s}\|_2^2 + \|\hat{t}\|_2^2) \quad (17)$$

where we set $\gamma := 1e-2$ for all of our experiments. We provide in-depth discussion regarding regularization and global scale g_s in the Supplemental 7.3.

Discussion. Unlike prior works that directly utilize the reprojection loss for self-supervised depth estimation [15, 16] or online adaptation [52, 65], our method incorporates the gradient of this loss as additional guidance within the diffusion sampling process. As Section 5.3 demonstrates, optimizing this reprojection loss in the raw pixel space is highly susceptible to noise and can lead to inferior depth maps. Consequently, previous methods have employed smoothness regularizers to mitigate this issue. By embedding the loss into the diffusion sampling framework, our diffusion model inherently acts as an implicit regularizer, effectively stabilizing the optimization and obviating the need for explicit smoothness constraints.

5. Experiments

In this section, we first describe our experimental settings at Section 5.1. Then, we showcase our experiment results in Section 5.2. Finally, we perform ablation study at Section 5.3.

5.1. Experiment setup

Implementation details. Our method is built upon Marigold [23]. Specifically, we employ their public pre-trained model and modify it with our guidance. While doing optimization, we completely freeze the trained weights of the model. The learning rate for optimizing parameters in Section 4.1 and the depth latents are set to $1e-2$. Following Marigold [23], we use an ensemble of 10 depth samples as our final prediction for computing all metrics. For fair comparison, all diffusion-based methods listed in Table 1 are also results of ensemble prediction. All experiments are conducted on a single NVIDIA RTX A6000 GPU.

Dataset. We perform zero-shot evaluation of our method on three datasets: KITTI-2015 [32], Booster [36] and Middlebury [44]. While KITTI-2015 and Middlebury are the two common outdoor and indoor benchmark for depth estimation, Booster is a more recent depth benchmark focusing on specular and transparent objects. We also sample a subset of multi-view depth dataset Tanks and Temples for evaluation of arbitrary two-view setting. More detail is shown at Supplemental 10.

Evaluation metrics. Following previous works [14, 23], we conduct zero-shot metric depth estimation by measuring three metrics including mean absolute error (AbsRel), $\delta 1$ accuracy. Additionally, we measure the root mean square error (RMSE) $\frac{1}{M} \sum_{i=1}^M \|d_i^{gt} - d_i^{pred}\|_2^2$, where M

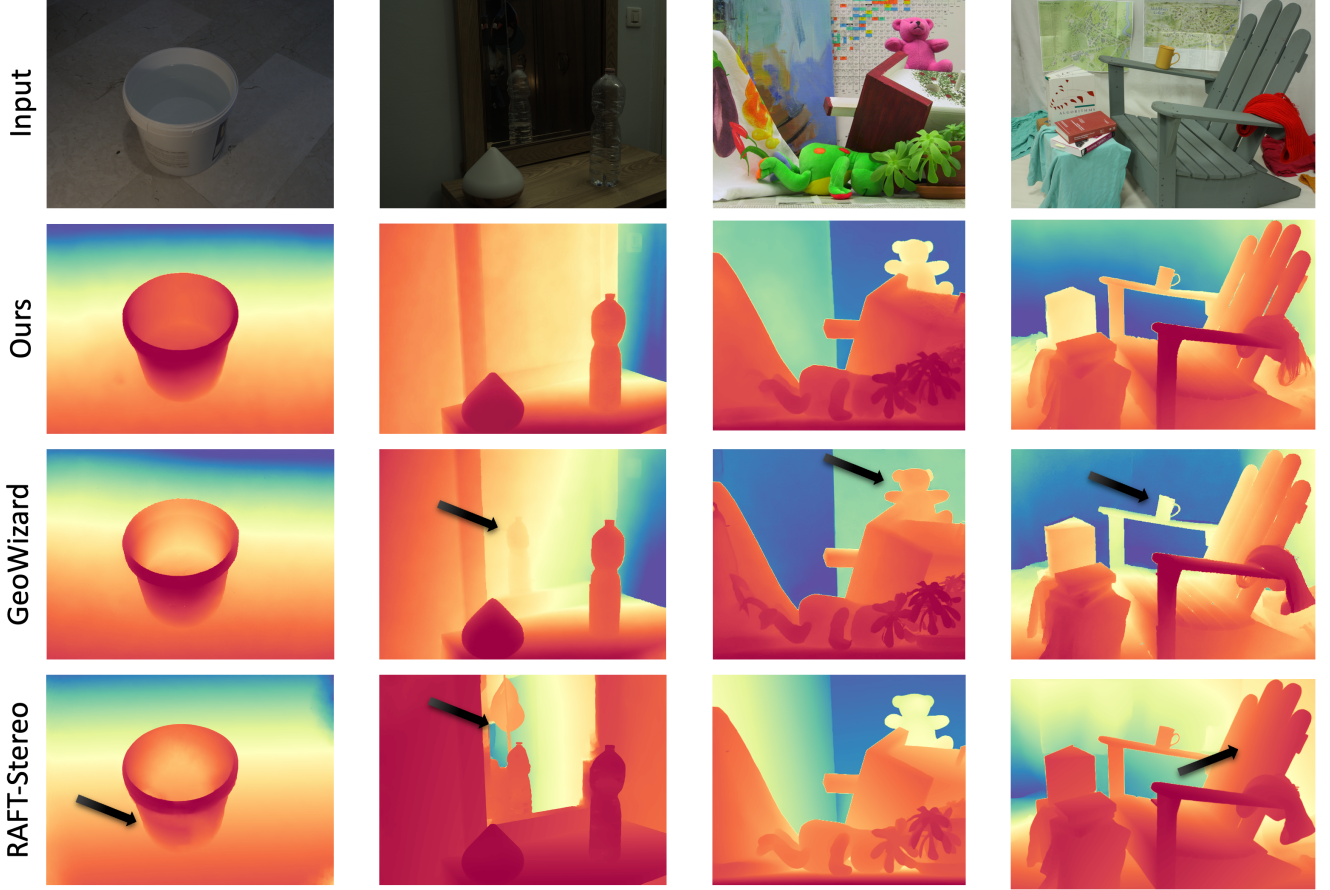


Figure 4. **Qualitative results.** Our method showcases superior depth prediction compared to competing methods. For instance, in the challenging Mirror scene (second column), our model accurately predicts the reflective surface, outperforming both GeoWizard [14] and RAFT-Stereo, which struggle in such cases. Additionally, our approach preserves finer details (last column), showcasing the effectiveness of our proposed geometric guidance.

denotes the number of samples, and d^{gt}, d^{pred} represent ground truth depth and predicted depth. To comprehensively demonstrate our method’s effectiveness, we report results for both predicted affine-invariant depth map and our metric depth maps. For affine-invariant depth map, we use the common protocol employed in previous works [14, 23] that align predicted depth map to ground truth via least-squares fitting. Conversely, our metric depth map is directly compare with ground truth depth without any alignment.

Competing methods. Our method utilizes a stereo pair as input while leveraging pretrained monocular priors without any stereo-specific training. To the best of our knowledge, this represents the first training-free approach that adapts monocular priors for stereo depth estimation. Consequently, there are no direct competing methods in the same setting. We therefore conduct comprehensive comparisons across related domains to contextualize our contributions. Since our method is based on monocular priors, we compare against recent state-of-the-art affine-invariant

MDE methods including Marigold [23], GeoWizard [14], and MiDas [4]. Following their evaluation protocol, we align their predicted depths with ground truth using least squares fitting [14, 23, 37]. As our method also recovers metric depth, we compare with leading metric depth estimation approaches: ZoeDepth [3], UniDepth [35], and DepthPro [5]. In this setting, we directly compare raw metric depth outputs without any alignment. Given our use of stereo pairs as input, we additionally benchmark against stereo matching-based methods including RAFT-Stereo [30] and IGEV-Stereo [58]. While these methods primarily predict disparity rather than depth, the comparison remains relevant as disparity ($disp$) can be directly converted to depth (d) given the baseline (b) and camera focal length (f): $d = bf / disp$.

5.2. Experimental results

5.2.1. Comparison with Monocular Depth Estimation.

Quantitative results. We present our quantitative results in Table 1. Our method demonstrates superior performance

Table 1. **Quantitative comparison with monocular depth methods on KITTI-Stereo, Booster, and Middlebury Datasets.** Our method outperforms existing approaches in both aligned and non-aligned settings. “GT-Aligned” indicates whether predictions are affine-invariant and require alignment with ground truth. Color coding: best, second-best, and third-best results. For metrics, $\downarrow$ indicates lower is better while $\uparrow$ indicates higher is better. We denote $\dagger$ as our result for metric depth, and the other result is affine-invariant depth.

Method	GT-Aligned	KITTI-Stereo			Booster			Middlebury		
		AbsRel $\downarrow$	$\delta 1\uparrow$	RMSE $\downarrow$	AbsRel $\downarrow$	$\delta 1\uparrow$	RMSE $\downarrow$	AbsRel $\downarrow$	$\delta 1\uparrow$	RMSE $\downarrow$
MiDas [4]	✓	0.63	0.24	11.72	0.18	0.71	0.23	0.22	0.72	2.09
Marigold [23]	✓	0.13	0.85	4.81	0.04	0.98	0.06	0.14	0.83	1.61
GeoWizard [14]	✓	0.18	0.75	5.7	0.04	0.96	0.08	0.15	0.81	1.55
ZoeDepth [3]	✗	0.74	\	16.26	7.37	\	7.72	0.60	\	6.04
UniDepth [35]	✗	0.19	0.86	4.18	4.55	\	4.91	0.61	\	6.03
DepthPro [5]	✗	0.16	0.81	4.43	0.34	0.51	0.67	0.54	0.02	5.60
Ours $\dagger$	✗	0.07	0.91	3.94	0.11	0.81	0.18	0.11	0.85	2.31
Ours	✓	0.09	0.91	3.72	0.04	0.98	0.06	0.11	0.87	1.41

in both aligned and non-aligned settings across all datasets. In the depth affine-invariant setting with ground truth alignment, our method significantly outperforms existing approaches across all metrics. We consistently surpass our baseline method Marigold [23] by a substantial margin. On both KITTI-Stereo and Middlebury datasets, our affine-invariant depth outperforms all competing methods including MiDas [4], Marigold [23], and GeoWizard [14]. In the non-aligned setting, our method also exhibits competitive performance across all datasets, even when compared with affine-invariant methods that require ground truth alignment. On KITTI-Stereo (an outdoor dataset), we achieve the lowest AbsRel of 0.07 and highest $\delta 1$ of 0.91. On Middlebury (an indoor dataset), our approach outperforms in AbsRel (0.11) and $\delta 1$ (0.85) metrics, falling short only in the RMSE metric. For Booster, a challenging dataset with non-Lambertian surfaces including transparent and reflective objects, we consistently outperform metric depth estimation methods such as ZoeDepth [3], UniDepth [35], and DepthPro [5]. This highlights our method’s robustness when handling challenging surface properties. It is notable that without retraining Marigold [23], our method outperforms diffusion-based monocular baselines in non-aligned settings while achieving superior performance with alignment. These results demonstrate our approach’s ability to produce accurate metric depth maps without requiring ground truth alignment, while still excelling when alignment is applied.

Qualitative results. We present our qualitative results in Figure 1 and Figure 4. Our method, which leverages both strong geometric guidance and pretrained diffusion priors, effectively captures fine-grained details while accurately representing transparent and specular objects. When compared to our baseline Marigold [23], our approach eliminates several erroneous artifacts due to our geometric guid-

Table 2. **Quantitative comparison with stereo depth on Booster Dataset.** Our methods performs comparably or better than stereo methods despite no explicit stereo training. While both aligned and non-aligned results are reported, we emphasize raw metric predictions (non-aligned), with affine-invariant results provided only for reference.

Method	GT-Aligned	Booster		
		AbsRel $\downarrow$	$\delta 1\uparrow$	RMSE $\downarrow$
RAFT-Stereo [30]	✗	5.4	\	18.89
IGEV-Stereo [58]	✗	0.10	0.94	0.31
Ours $\dagger$	✗	0.11	0.81	0.18
Ours	✓	0.04	0.98	0.06

ance (as shown in Figure 1). In comparison with GeoWizard [14], our method correctly handles reflective surfaces such as mirrors and captures more detailed depth information in indoor scenes, as illustrated in Figure 4. Although DepthPro [5] may look sharper in 2D visual results, our depth predictions are superior in terms of metric accuracy, as depth inherently represents 3D information. More visualization results are presented in Supplemental 12.

5.2.2. Comparison with Stereo Depth Estimation

Quantitative results. We present our quantitative result in stereo depth estimation setting at Table 2. It is worth noting that previous stereo methods such as RAFT-Stereo [30] and IGEV-Stereo [58] have been trained KITTI-Stereo dataset and are extensively tuned for Middlebury, yet remain unevaluated on the challenging Booster dataset. Therefore, for fair comparison with our method—as the pre-trained never encountered Booster during training—we conduct comparative analysis on this dataset. In non-alignment setting, we achieve on-par result compare to strong stereo method IGEV-Stereo in AbsRel and $\delta 1$ (0.11 and 0.81, respectively). Notably, we outperform both stereo methods

Table 3. **Quantitative comparison with stereo depth on Tank and Temple.** Our method excels in metric depth estimation, and on par with Dust3r when aligned with ground truth.

Method	GT-Aligned	Tanks and Temples		
		AbsRel↓	$\delta 1 \uparrow$	RMSE↓
Dust3r [54]	×	0.51	0.35	0.56
Ours [†]	×	0.47	0.38	0.62
Dust3r [54]	✓	0.07	0.93	0.17
Ours	✓	0.08	0.92	0.18

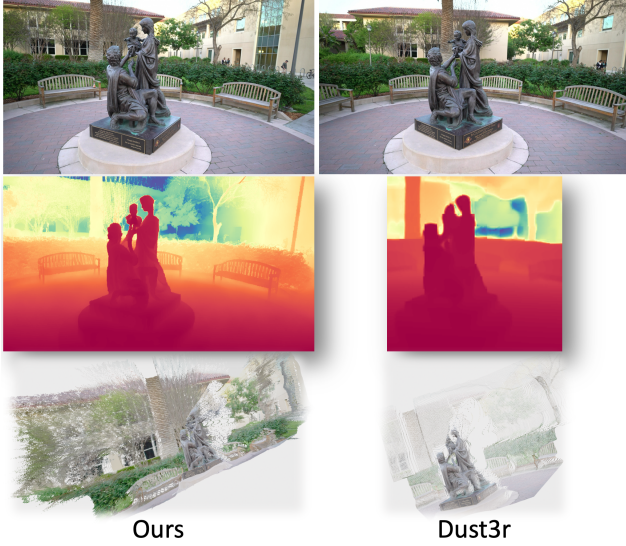


Figure 5. From top to bottom. **Top:** Two known-pose arbitrary images. **Middle:** Depth map predictions of our method and Dust3r [54]. **Bottom:** Our point cloud and Dust3r point cloud.

in the RMSE metric with a value of 0.18. These results demonstrate that our geometry-guided diffusion prior approach performs consistently on par or better with stereo method that has been exposed to stereo training data.

Qualitative results. As shown in Figure 1 and Figure 4, our method excels on cases that contain non-Lambertian surfaces such as transparent materials and reflective surfaces. We speculate that though stereo methods are extensively trained on stereo data, they heavily rely on cost volume-based approach, which inherently struggles with transparent or textureless surfaces [36]. Our method, on the other hand, inherits strong monocular priors from diffusion models guided with explicit geometry guidance during optimization, thus addressing the cost volume limitation. Furthermore, a strong geometry-guided monocular prior also help to achieve shaper depth in fine-grain areas (see Middlebury results in Figure 4).

Table 4. Ablation Study on KITTI-Stereo Dataset

Method	AbsRel↓	$\delta 1 \uparrow$
Learning scale and shift only	0.14	0.80
Our full model	0.07	0.91

5.2.3. Arbitrary two-view settings

As established in Section 4.2, our methodology extends beyond traditional stereo setups to general two-view configurations with known camera poses. While numerous multi-view stereo techniques can infer depth from multiple viewpoints [56, 59, 62], our work focuses specifically on the two-view paradigm. We compare our method with Dust3r [54], which is trained on two views to generate a point map from a single view. Quantitative results are presented in Table 3, with qualitative evaluation shown in Figure 5. Compared to Dust3r [54], our approach produces significantly more detailed depth maps with enhanced structural fidelity. Although Dust3r generates visually plausible point clouds, their reconstructions are limited to an unknown scale factor. In contrast, our method produces metric point clouds that accurately represent scene geometry at absolute scale, enabling precise spatial measurements and supporting reliable downstream applications.

5.3. Ablation studies

5.3.1. Non-optimality of the reprojection loss

We empirically investigated the noise sensitivity of reprojection loss discussed in Section 4. By directly optimizing metric depth without the diffusion framework—initializing with Marigold outputs and optimizing both depth and scale/shift parameters over 50 iterations—we observe significantly noisy results as shown in Figure 6. In contrast, our diffusion-based approach preserves fine details while avoiding noise artifacts. We hypothesize that the diffusion sampling process inherently regularizes the optimization toward the true depth distribution, eliminating the need for explicit smoothness regularizers [15, 16].

5.3.2. Effectiveness of the reprojection loss

To investigate whether reprojection loss primarily enhances depth quality or merely calibrates scale/shift parameters, we conducted an ablation study restricting this loss to optimize only scale and shift without influencing diffusion latents. As shown in Table 4, experiments on KITTI-Stereo demonstrate significant performance degradation under this configuration, confirming the crucial role of reprojection loss guidance in refining the diffusion process for accurate metric depth estimation.

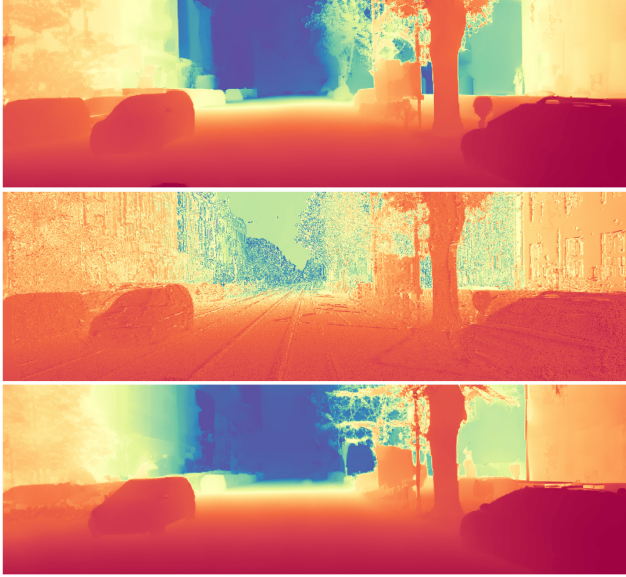


Figure 6. **Ablation.** From top to bottom: *Marigold prediction*, *optimization with reprojection loss only*, and *ours combining diffusion with reprojection loss guidance*. Reprojection loss optimization leads to noisy and suboptimal depth, while using it as the guidance for diffusion model helps improve the results.

6. Conclusion and Limitation

In this work, we introduced a novel framework that extends diffusion-based monocular depth estimation (DB-MDE) models to metric depth prediction by incorporating stereo settings and an inverse problem (IP) approach. By leveraging pretrained latent diffusion models (LDMs) with stereo geometric guidance, our method effectively addresses scale and shift ambiguities inherent in monocular depth estimation. Extensive experiments demonstrate its robustness across diverse environments, including indoor, outdoor, and challenging specular scenes, all without requiring domain-specific retraining.

Despite its strengths, our approach has certain limitations. First, it relies on pretrained monocular depth estimation models, meaning the quality of depth predictions is dependent on the strength of the prior. A more expressive or robust MDE model could further enhance performance. Second, like other DB-MDE approaches, our method incurs slow inference times due to the iterative nature of diffusion-based sampling. Future work could explore accelerated sampling techniques or lighter-weight diffusion architectures to improve efficiency while maintaining accuracy.

GeoDiff: Geometry-Guided Diffusion for Metric Depth Estimation

Supplementary Material

In this supplementary material, we first provide additional derivations and insights in Section 7. We then present our method through pseudocode in Section 8, followed by implementation details in Section 9. Dataset specifications are described in Section 10, while limitations and future work are discussed in Section 11. Finally, additional experimental results are presented in Section 12.

7. Detailed method

7.1. Conditional Score Derivation

In this section, we provide the complete derivation of our conditional score. Applying Bayes' theorem, the score function of the conditional distribution can be expressed as:

$$\begin{aligned} \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t | \mathbf{y}_1, \mathbf{y}_2) &= \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t | \mathbf{y}_1) \\ &\quad + \nabla_{\mathbf{z}_t} \log p(\mathbf{y}_2 | \mathbf{z}_t, \mathbf{y}_1) \end{aligned} \quad (18)$$

Under mild assumptions [7], and a decoder $\mathbf{D}$ that maps the latent back to the image space, we can approximate this score function using:

$$\begin{aligned} \nabla_{\mathbf{z}_t} \log p(\mathbf{y}_2 | \mathbf{z}_t, \mathbf{y}_1) &\approx \nabla_{\mathbf{z}_t} \log p(\mathbf{y}_2 | \hat{\mathbf{z}}_0, \mathbf{y}_1) \\ &\approx \nabla_{\mathbf{z}_t} \log p(\mathbf{y}_2 | \mathbf{D}(\hat{\mathbf{z}}_0), \mathbf{y}_1) \end{aligned} \quad (19)$$

where $\hat{\mathbf{z}}_0$ is estimated using Tweedie's formula [40]:

$$\hat{\mathbf{z}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{z}_t + \sqrt{1 - \bar{\alpha}_t} \mathbf{s}_\theta(\mathbf{z}_t, t, \mathbf{y}_1)) \quad (20)$$

Leveraging the Gaussian noise model assumption in Equation 11, we get:

$$\nabla_{\mathbf{z}_t} \log p(\mathbf{y}_2 | \mathbf{z}_t, \mathbf{y}_1) \simeq -\frac{1}{\sigma^2} \nabla_{\mathbf{z}_t} \|\mathbf{y}_2 - \mathbf{P}_{\mathbf{y}_1 \rightarrow \mathbf{y}_2}(\tilde{\mathbf{x}}_0, \mathbf{y}_1)\|_2^2 \quad (21)$$

in which $\tilde{\mathbf{x}}_0$ is the metric depth calculated following the Equation 13.

Thus, the final conditional score function is:

$$\begin{aligned} \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t | \mathbf{y}_1, \mathbf{y}_2) &\approx \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t | \mathbf{y}_1) \\ &\quad - \lambda \nabla_{\mathbf{z}_t} \|\mathbf{y}_2 - \mathbf{P}_{\mathbf{y}_1 \rightarrow \mathbf{y}_2}(\tilde{\mathbf{x}}_0, \mathbf{y}_1)\|_2^2 \end{aligned} \quad (22)$$

7.2. Differentiable warping

As discussed in Section 3.3 and Section 4.2, we use differentiable warping to render novel view given predicted depth. Then, our method leverages given RGB input image to calculate photometric loss as guidance (see Equation 14) for diffusion process. There are two design choices

for warping operator, which are forward warping operator $\mathbf{P}_{\mathbf{y}_1 \rightarrow \mathbf{y}_2}(\mathbf{x}_1, \mathbf{y}_1)$, which projects $\mathbf{y}_1$ onto $\mathbf{y}_2$; and the backward warping operator $\mathbf{P}_{\mathbf{y}_2 \rightarrow \mathbf{y}_1}(\mathbf{x}_1, \mathbf{y}_2)$. If one opts to use forward warping as renderer, $\mathcal{L}_{geo}$ in Equation 14 has the following form:

$$\begin{aligned} \mathcal{L}_{geo}^{forward} &= \eta(1 - \text{SSIM}(\mathbf{y}_2, \mathbf{P}_{\mathbf{y}_1 \rightarrow \mathbf{y}_2}(\tilde{\mathbf{x}}_0, \mathbf{y}_1)))/2 \\ &\quad + (1 - \eta) \|\mathbf{y}_2 - \mathbf{P}_{\mathbf{y}_1 \rightarrow \mathbf{y}_2}(\tilde{\mathbf{x}}_0, \mathbf{y}_1)\|_1 \end{aligned} \quad (23)$$

otherwise, the backward warping could also be used with the form:

$$\begin{aligned} \mathcal{L}_{geo}^{backward} &= \eta(1 - \text{SSIM}(\mathbf{y}_1, \mathbf{P}_{\mathbf{y}_2 \rightarrow \mathbf{y}_1}(\tilde{\mathbf{x}}_0, \mathbf{y}_2)))/2 \\ &\quad + (1 - \eta) \|\mathbf{y}_1 - \mathbf{P}_{\mathbf{y}_2 \rightarrow \mathbf{y}_1}(\tilde{\mathbf{x}}_0, \mathbf{y}_2)\|_1 \end{aligned} \quad (24)$$

Now, we discuss the design choice of the two options. Given a source image ($\mathbf{y}_1$), target image ($\mathbf{y}_2$), intrinsic camera matrices K_1, K_2 , and camera-to-world extrinsic matrices E_1, E_2 for the source and target views respectively, we establish a generalized framework for our method. We explicitly represent both camera extrinsics to handle arbitrary camera configurations rather than just using a single relative transformation $T_{1 \rightarrow 2}$ as in Equation 9. In the specific case of calibrated stereo pairs captured simultaneously by a binocular rig, the transformation simplifies to a pure translation. However, for arbitrarily captured image pairs, the complete extrinsic matrices are necessary to accurately transform points between the two coordinate systems.

Forward warping maps pixels from a source image ($\mathbf{y}_1$) to positions in target image ($\mathbf{y}_2$). The target coordinate is formulated as:

$$\mathbf{c}_2 \sim K_2 E_2^{-1} E_1 \mathbf{x}_1(\mathbf{c}_1) K_1^{-1} \mathbf{c}_1 \quad (25)$$

where $\mathbf{c}_1$ and $\mathbf{c}_2$ denote the homogeneous pixel coordinates in $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively, and $\mathbf{x}_1(\mathbf{c}_1)$ represents the depth at pixel $\mathbf{c}_1$ in $\mathbf{y}_1$. After getting the corresponding pixel coordinates, we can "splat" each source pixel to its corresponding location in target view. However, there are a few implementation challenges. A fundamental issue is that some target pixels might not receive any values, creating holes in the warped image. These voids occur due to disocclusions (regions visible in the target view but occluded in the source view) and sampling disparities (discrete source pixels mapping to non-integer target coordinates with gaps between them). Addressing these holes requires complex post-processing techniques such as depth-aware inpainting or multi-scale filtering. The non-integer mapping of source

pixels to target coordinates further introduces discretization errors and potential aliasing, requiring appropriate interpolation strategies. Depth map inaccuracies are particularly problematic at discontinuities, where slight errors can significantly distort the warped result, making it even more difficult to apply to our framework. From a computational perspective, the unpredictable memory access patterns inherent in forward warping present optimization difficulties, particularly for parallel processing implementations.

Backward warping pulls back pixels from target image (y_2) to source image (y_1). It is worth noting that our backward warping is different from previous works [15, 28], where they perform backward warping from source to target given target depth. On the other hand, we warp from target back to source using source depth. Specifically, our backward warping is formulated as:

$$c_2 \sim K_2 E_2 E_1^{-1} x_1(c_1) K_1^{-1} c_1 \quad (26)$$

After computing the corresponding pixel coordinates, we sample pixel colors from the target image at these new coordinates. Since these coordinates are generally non-integer, we employ bilinear interpolation for color sampling. This approach inherently avoids the hole artifacts characteristic of forward warping methods. For this reason, we adopt backward warping as our rendering technique throughout this work.

Discussion. We explored several alternative techniques that ultimately proved suboptimal. Initial experiments with point cloud rasterization from Pytorch3D [38] revealed high sensitivity to point diameter and opacity parameters, resulting in rendering artifacts including holes and visible disk-like structures. Similarly, we attempted to initialize the point cloud as 3D Gaussians (3DGS) to leverage recent differentiable Gaussian rasterization techniques [25]. However, the 3DGS renderer introduces an excessive number of parameters to optimize, which proved inefficient during the limited sampling steps of our diffusion process.

Left-right consistency. While left-right consistency checks are commonly employed in stereo methods [15], we deliberately omit this approach in our framework. Unlike learning-based methods that can infer depth of both left and right from a single image, our optimization-based technique would require running the depth prediction process twice—once for each view—effectively doubling the computational cost. Therefore, in this work, we demonstrate our method’s efficacy by optimizing the photometric loss using only a single reference view, achieving a favorable balance between accuracy and computational efficiency.

7.3. Regularization

As described in Section 4, we stabilize the optimization process by introducing a global scale g_s and applying L_2 regularization to the scale $\hat{s}$ and shift $\hat{t}$ parameters. To illustrate the design of these parameters, we present a toy example. Given a predicted relative depth map $\tilde{x}_0^{rel}$ (normalized to the range $[0, 1]$), we incrementally increase the global scale g_s to compute the scaled depth $\tilde{x}_0^{scale}$ such that $\tilde{x}_0^{scale} = g_s \cdot \tilde{x}_0^{rel}$. For each scale, we evaluate the Absolute Relative (AbsRel) metric (where lower values are better) using the scaled depth $\tilde{x}_0^{scale}$ and the ground truth depth. Additionally, using the left image, right image, and the predicted scaled depth $\tilde{x}_0^{scale}$, we compute the reprojection loss $\mathcal{L}_{geo}$, as defined in Equation 14, for each scale.

As shown in Figure 10, increasing the depth scale improves the resemblance of the re-rendered image to the left image. This occurs because closer depths result in larger disparities between the source and target viewpoints, requiring more significant transformations to align the images. Such large transformations often cause distortions, stretching, or undersampling in areas lacking sufficient source information, degrading the quality of the re-rendered image. In contrast, at greater depths, disparities between the viewpoints are smaller, leading to less dramatic transformations. These smaller adjustments maintain spatial coherence more effectively and reduce interpolation artifacts, producing sharper and more accurate re-rendered images.

Figures 7 and 8 demonstrate this pattern as the global scale g_s increases from 1 to 50. However, further increasing g_s , while improving the re-rendered image quality and enhancing $\mathcal{L}_{geo}$, leads to worse AbsRel metrics. This indicates that the depth scale $\tilde{x}_0^{scale}$ deviates from the ground truth depth. This behavior underscores the strong geometric guidance provided by $\mathcal{L}_{geo}$ for the diffusion model during sampling.

To balance these considerations, we pre-select g_s based on the geometric loss. Specifically, we search for g_s within a predefined depth range and define the optimal scale as $g_s^* := \arg \min_{g_s} \mathcal{L}_{geo}$. Our approach can be viewed as a variant of the traditional Plane Sweep Volume technique [9], commonly used in stereo vision. Unlike conventional methods, our approach leverages the predicted relative depth to identify the correct depth scale, which is then applied uniformly to all pixels.

Finally, we apply L_2 regularization to the scale and shift parameters of $\tilde{x}_0^{rel}$ to counteract the tendency of the optimization process to inflate these parameters, which can lead to incorrect metric depth predictions.

8. Algorithm

We provide detail pseudo algorithm for our method at Algorithm 1. To avoid confusion, note that while learnable scale

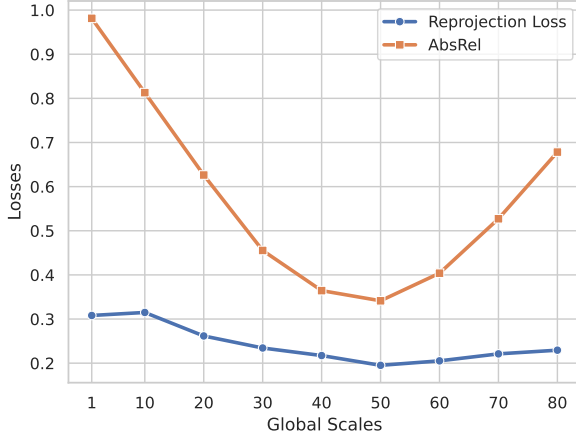


Figure 7. Global scale up to 80

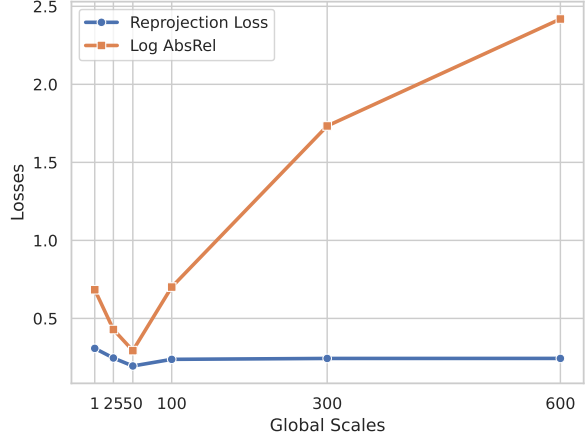


Figure 8. Global scale up to 600

Figure 9. We gradually increase the global scale g_s and observe a strong correlation between the reprojection loss and the AbsRel metric. For this example, the AbsRel reaches its minimum at a global scale of 50. However, beyond 50, the AbsRel significantly increases, while the reprojection loss shows little change, deviating from the pattern.

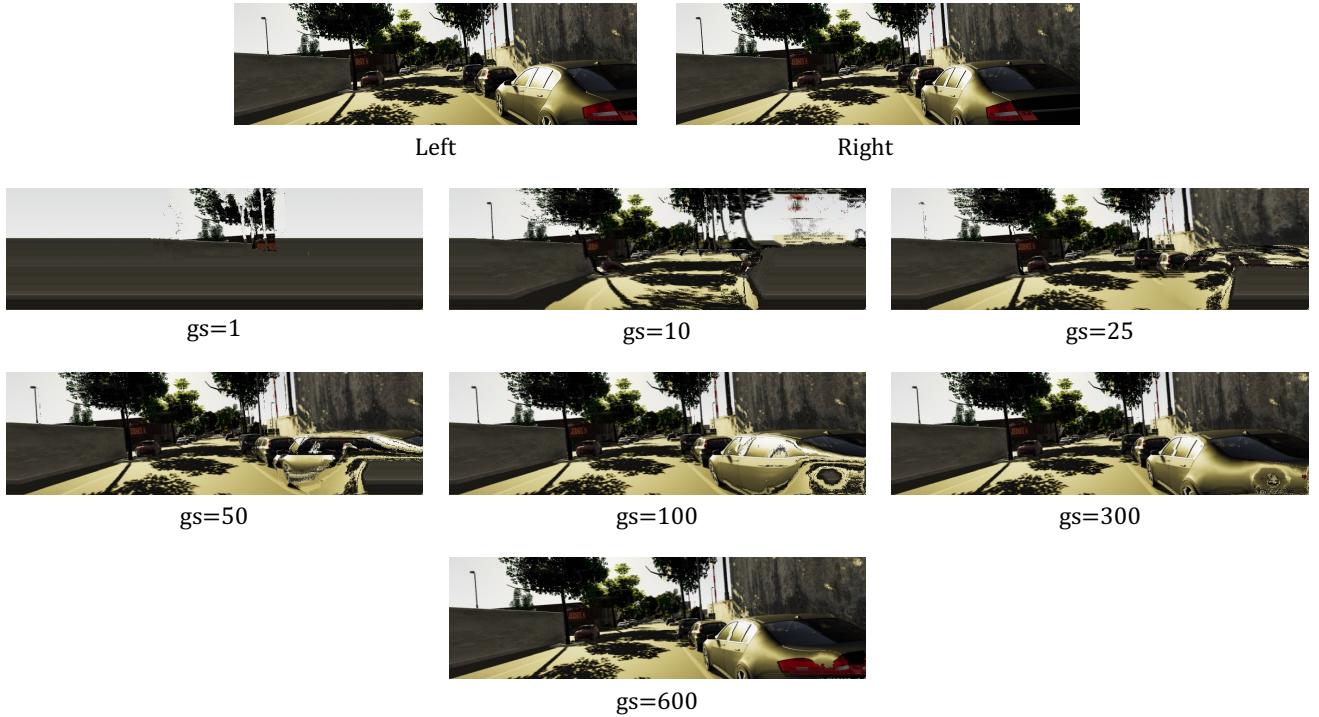


Figure 10. We gradually increase the global scale g_s and re-render the left image using the right image and $\hat{x}_0^{scale}$ (see Section 7.3). While greater depths result in higher-quality re-rendered images, this does not necessarily correspond to more accurate predicted depths.

and shift are denoted as $(\hat{s}, \hat{t})$, score function and time step of diffusion are denoted as (s, t) , respectively.

9. Implementation details

Geometric optimization steps. Our method employs a test-time optimization approach. While multiple gradient updates could theoretically be performed during sampling

Algorithm 1 Geometric-Guided Diffusion for Metric Depth Estimation

Require: Stereo images y_1, y_2 , camera intrinsics and extrinsics, pretrained diffusion model $s_\theta(z_t, t, y_1)$

Initialize learnable scale $\hat{s}$ and shift $\hat{t}$

Initialize random noise $z_T \sim \mathcal{N}(0, I)$.

for $t = T - 1$ to 0 **do**

$\hat{s}_{t+1} = s_\theta(z_t, t, y_1)$

▷ Compute the score

$\hat{z}_0 = \frac{1}{\sqrt{\alpha_t}} (z_t + \sqrt{1 - \alpha_t} s_\theta(z_t, t, y_1))$

▷ Compute relative depth using Tweedie’s formula

$\tilde{x}_0 = \text{softplus}(\hat{s}) \cdot D(z_0) + \text{softplus}(\hat{t})$

▷ Convert relative depth to metric scale

Compute $\mathcal{L}_{geo}$ following Eq. 14

$\hat{s} \leftarrow \hat{s} - \lambda_{\hat{s}} \nabla_{\hat{s}} \mathcal{L}_{geo}$

▷ Gradient update for $\hat{s}$

$\hat{t} \leftarrow \hat{t} - \lambda_{\hat{t}} \nabla_{\hat{t}} \mathcal{L}_{geo}$

▷ Gradient update for $\hat{t}$

$z_{t-1} = \sqrt{\alpha_{t-1}} \hat{z}_0 + \sqrt{1 - \alpha_{t-1}} s_\theta(z_t, t, y_1) - \lambda \nabla_{z_t} \mathcal{L}_{geo}$

▷ Perform DDIM step with geometric guidance

end for

Output: Estimated metric depth map $\tilde{x}_0$

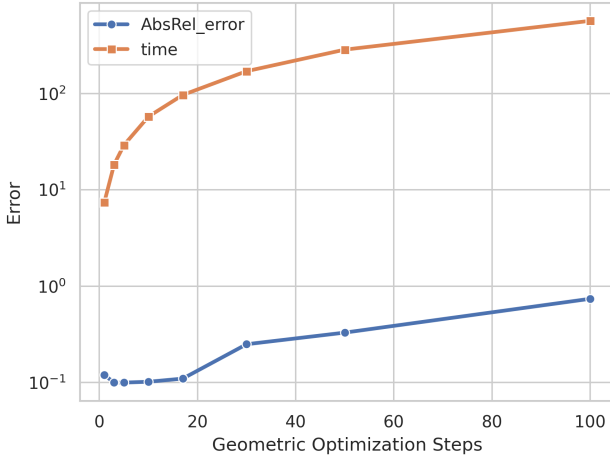


Figure 11. We increase the number of times to update inside one sampling step. We observe that it not only does not improve result, but also very time consuming to run.

to minimize the geometric loss, our experiments in Figure 11 demonstrate that only a limited number of gradient steps are beneficial. Consequently, we implement a single gradient update per sampling step in this work. Empirical observations indicate that increasing the number of gradient updates not only fails to improve performance but also significantly increases computational time.

Inference time. Inference time for a single sample using our approach is approximately 7 seconds on an RTX A6000 GPU with images of 768 pixels in dimension. This measurement excludes data preparation time, which varies across datasets.

Run time comparison. Our method requires test-time optimization but maintains computational efficiency com-

parable to the baseline Marigold [23], adding only seconds per image to processing time. This efficiency stems from our implementation of single-step gradient updates with minimal learnable parameters. Additionally, our depth warping-based rendering technique is both fast and fully differentiable. Consequently, despite achieving superior results, our approach does not significantly increase computational overhead compared to Marigold.

10. Dataset details

Training data. Our proposed method is a test time optimization-based, so we do not require any training sample. For details about training dataset of our baseline method, we refer reader to Marigold [23].

Evaluation data. We evaluate our proposed approach on four distinct datasets. The KITTI-2015 dataset comprises 200 stereo pairs depicting outdoor scenes. The Middlebury dataset contains 15 stereo pairs predominantly featuring indoor environments. The Booster dataset includes 228 stereo pairs with challenging non-Lambertian surfaces. For the Tanks and Temples dataset, we randomly sampled 116 image pairs from a multi-view dataset spanning four scenes.

11. Limitation

Since our method is based on diffusion sampling process, it is not suitable for real time application. Additionally, since we employ depth warping as a rendering technique and utilize photometric loss as an optimization objective, our approach exhibits sensitivity to significant illumination variations between stereo images. Potential solutions include applying color correction prior to image rendering or implementing left-right consistency as described in Section 7.2. We defer these improvements to future work.

12. Additional qualitative results

Additional qualitative results are presented in Figure 12, Figure 13, Figure 14.

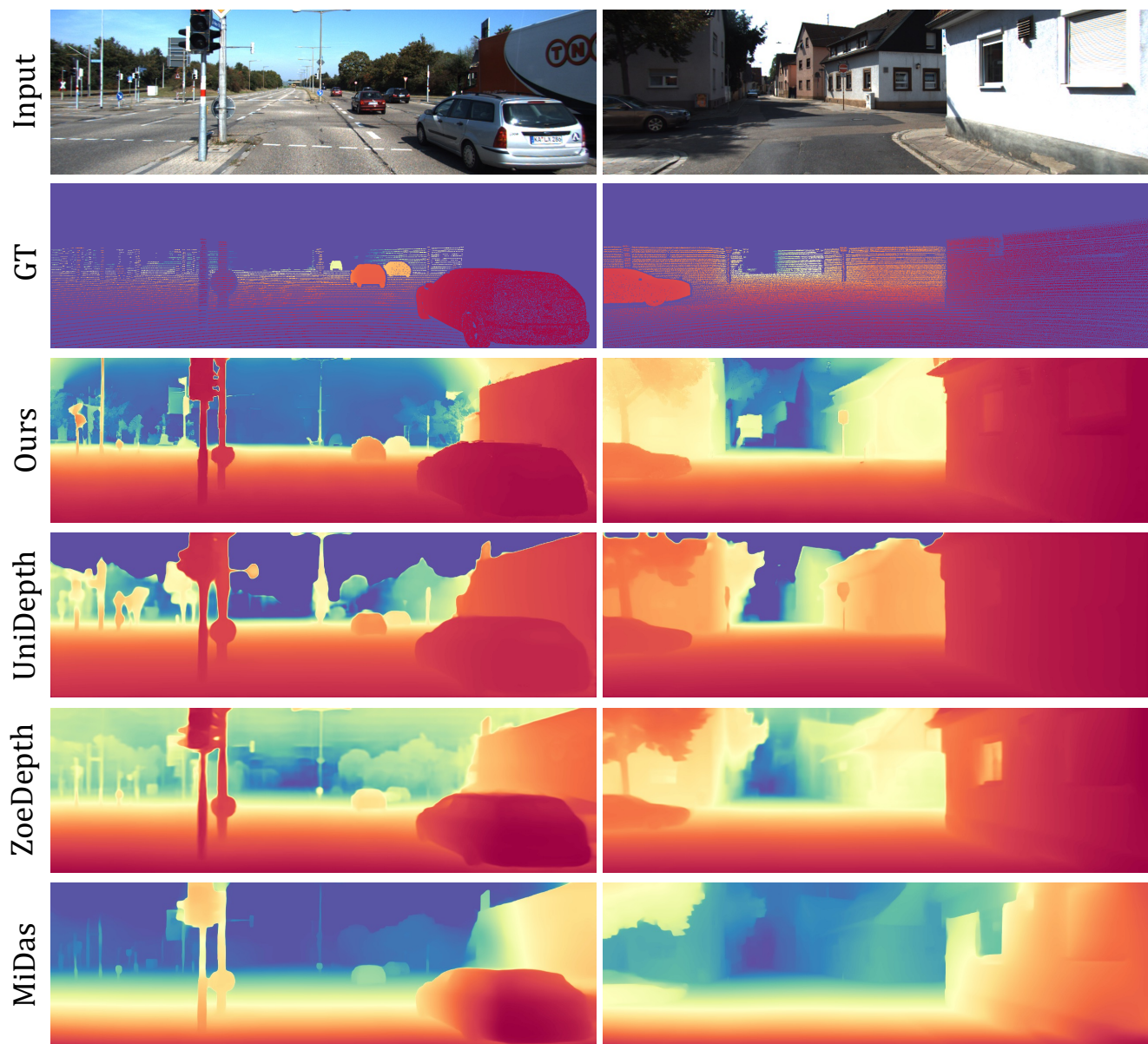
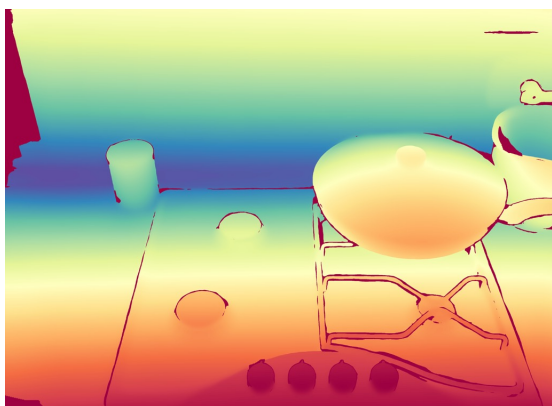


Figure 12. **Additional qualitative results.** Our method demonstrates superior depth quality in metric depth estimation, particularly in high-depth-range scenarios such as those found in the KITTI-2015 dataset [32], outperforming competing approaches.

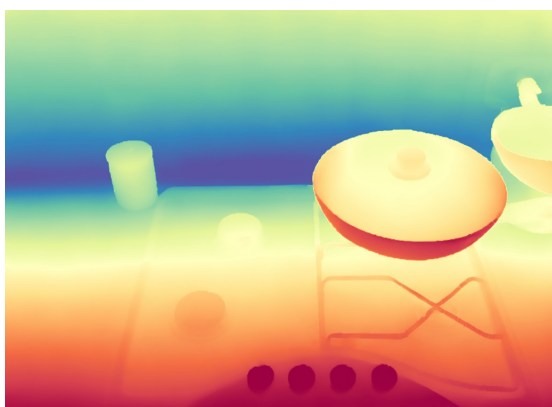
Input



GT



Ours



UniDepth

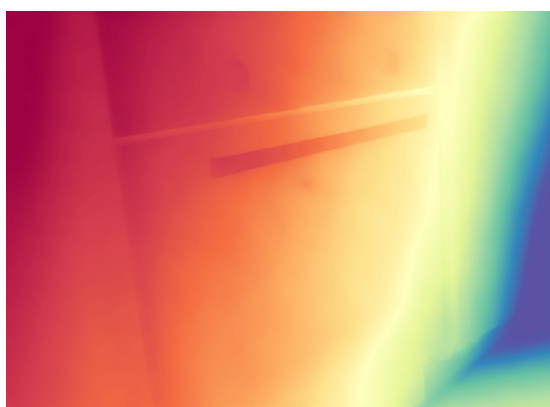
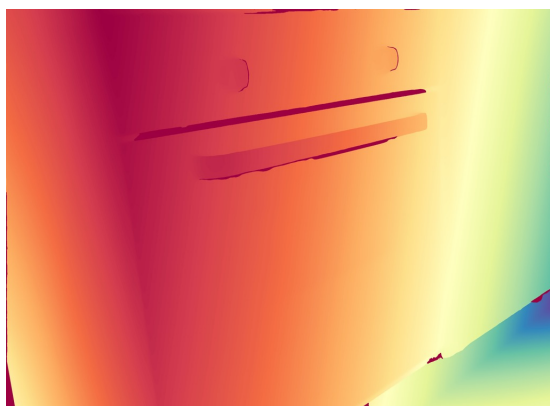
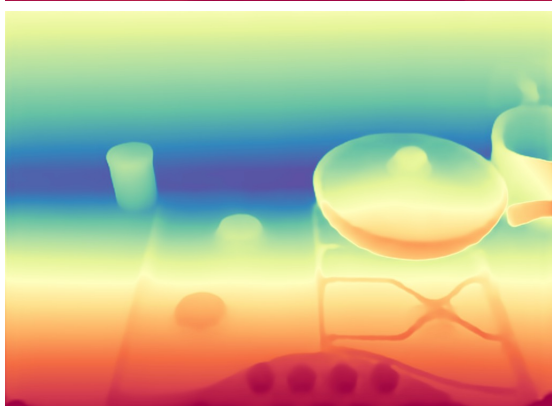


Figure 13. Additional qualitative results on Booster dataset [36]

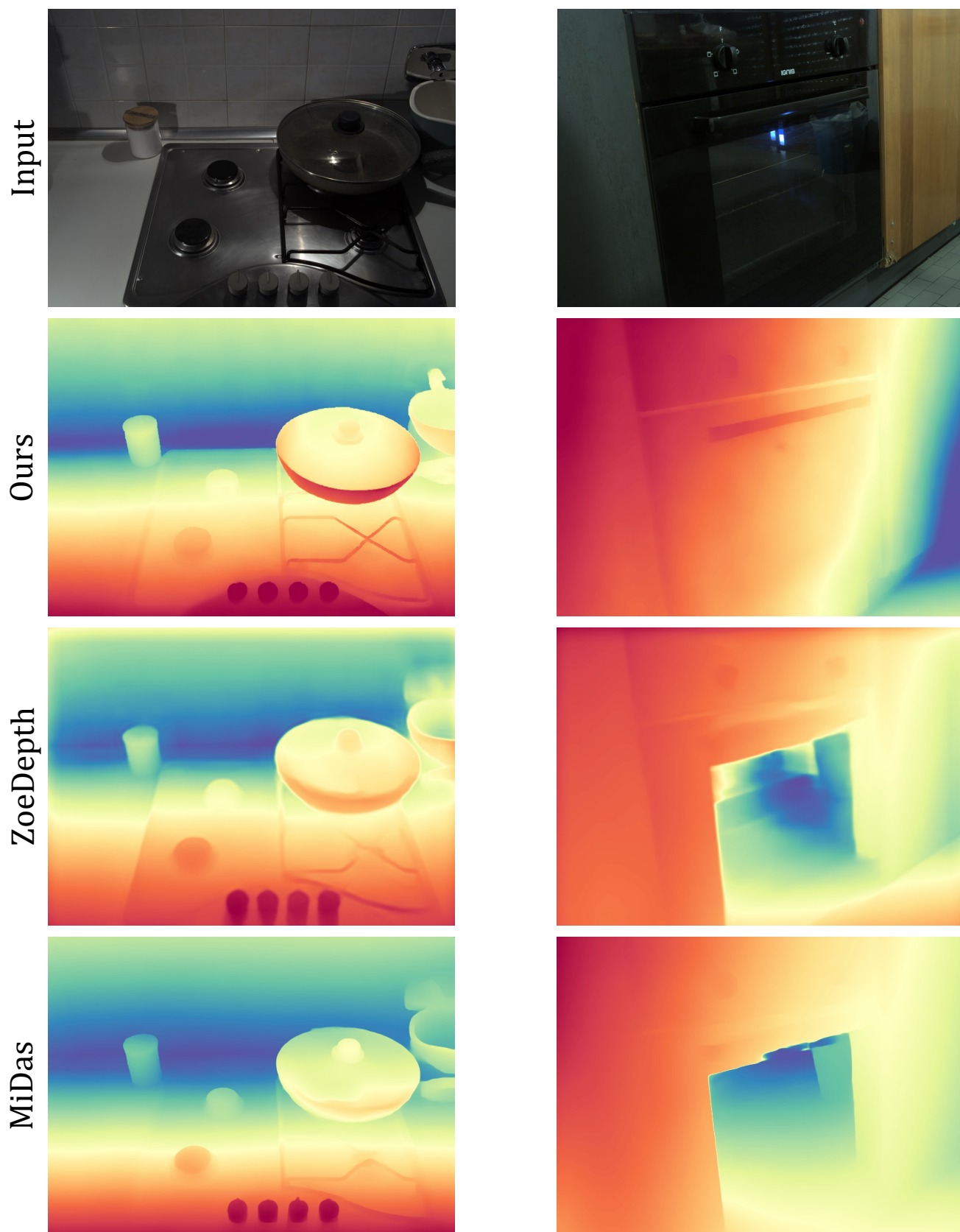


Figure 14. Additional qualitative results on Booster dataset [36]

References

- [1] Manyá V Afonso, José M Bioucas-Dias, and Mário AT Figueiredo. An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems. *IEEE transactions on image processing*, 20(3):681–695, 2010. 3
- [2] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009. 3
- [3] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 6, 7
- [4] Reiner Birkel, Diana Wofk, and Matthias Müller. Midas v3.1—a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 1, 6, 7
- [5] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 1, 2, 6, 7
- [6] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 2
- [7] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 2, 3, 4, 1
- [8] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022. 2, 3, 4
- [9] Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE computer society conference on computer vision and pattern recognition*, pages 358–363. Ieee, 1996. 2
- [10] Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024. 2
- [11] Yiqun Duan, Xianda Guo, and Zheng Zhu. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023. 2
- [12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 1
- [13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 1
- [14] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. *arXiv preprint arXiv:2403.12013*, 2024. 2, 5, 6, 7
- [15] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 3, 5, 8, 2
- [16] Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 3, 5, 8
- [17] Ming Gui, Johannes S Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024. 2
- [18] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3273–3282, 2019. 2
- [19] Kun Han, Shanlin Sun, Thanh-Tung Le, Xiangyi Yan, Haoyu Ma, Chenyu You, and Xiaohui Xie. Hybrid neural diffeomorphic flow for shape representation and generation via triplane. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7707–7717, 2024. 2
- [20] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Liu, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024. 2
- [21] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21741–21752, 2023. 2
- [22] Bahjat Kavar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 2
- [23] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 1, 2, 5, 6, 7, 4
- [24] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 2
- [25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2
- [26] Tung Le, Khai Nguyen, Shanlin Sun, Kun Han, Nhat Ho, and Xiaohui Xie. Diffeomorphic mesh deformation via efficient

- optimal transport for cortical surface reconstruction. *arXiv preprint arXiv:2305.17555*, 2023. 2
- [27] Tung Le, Khai Nguyen, Shanlin Sun, Nhat Ho, and Xiaohui Xie. Integrating efficient optimal transport and functional maps for unsupervised shape correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23188–23198, 2024. 2
- [28] Daniel Lichy, Hang Su, Abhishek Badki, Jan Kautz, and Orazio Gallo. nvtorchcam: An open-source library for camera-agnostic differentiable geometric vision. *arXiv preprint arXiv:2410.12074*, 2024. 2
- [29] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 1
- [30] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 2, 6, 7
- [31] Wei Liu, Xin Xia, Lu Xiong, Yishi Lu, Letian Gao, and Zhuoping Yu. Automated vehicle sideslip angle estimation considering signal measurement characteristic. *IEEE Sensors Journal*, 21(19):21675–21687, 2021. 3
- [32] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, 2:427–434, 2015. 5, 6
- [33] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1
- [34] Suraj Patni, Aradhye Agarwal, and Chetan Arora. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28285–28295, 2024. 2
- [35] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 6, 7
- [36] Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Booster: A benchmark for depth from images of specular and transparent surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):85–102, 2024. 1, 5, 8, 7
- [37] 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):1623–1637, 2020. 6
- [38] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 2
- [39] Saiprasad Ravishankar, Jong Chul Ye, and Jeffrey A Fessler. Image reconstruction: From sparsity to data-adaptive methods and machine learning. *Proceedings of the IEEE*, 108(1): 86–109, 2019. 3
- [40] Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 388–394. Springer, 1992. 3, 1
- [41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [42] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. In *arXiv*, 2023. 1, 2
- [43] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [44] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 31–42. Springer, 2014. 5
- [45] Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. *arXiv preprint arXiv:2307.08123*, 2023. 2, 3
- [46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4
- [47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3
- [48] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021. 3, 4
- [49] Paul Suetens. *Fundamentals of medical imaging*. Cambridge university press, 2017. 3
- [50] Shanlin Sun, Thanh-Tung Le, Chenyu You, Hao Tang, Kun Han, Haoyu Ma, Deyang Kong, Xiangyi Yan, and Xiaohui Xie. Hybrid-csr: Coupling explicit and implicit shape representation for cortical surface reconstruction. *arXiv preprint arXiv:2307.12299*, 2023. 2
- [51] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2
- [52] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip HS Torr. Learning to adapt for stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9661–9670, 2019. 3, 5

- [53] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. [3](#)
- [54] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. [5](#), [8](#)
- [55] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudolidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8445–8453, 2019. [1](#)
- [56] Jiang Wu, Rui Li, Haoifei Xu, Wenxun Zhao, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Gomvs: Geometrically consistent cost aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20207–20216, 2024. [8](#)
- [57] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12981–12990, 2022. [2](#)
- [58] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. [1](#), [2](#), [6](#), [7](#)
- [59] Hongbin Xu, Weitao Chen, Baigui Sun, Xuansong Xie, and Wenxiong Kang. Robustmvs: Single domain generalized deep multi-view stereo. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. [8](#)
- [60] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. [1](#), [2](#)
- [61] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. [1](#), [2](#)
- [62] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. [8](#)
- [63] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. [1](#)
- [64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. [1](#)
- [65] Zhenyu Zhang, Stéphane Lathuiliere, Andrea Pilzer, Nicu Sebe, Elisa Ricci, and Jian Yang. Online adaptation through meta-learning for stereo depth estimation. *arXiv preprint arXiv:1904.08462*, 2019. [3](#), [5](#)