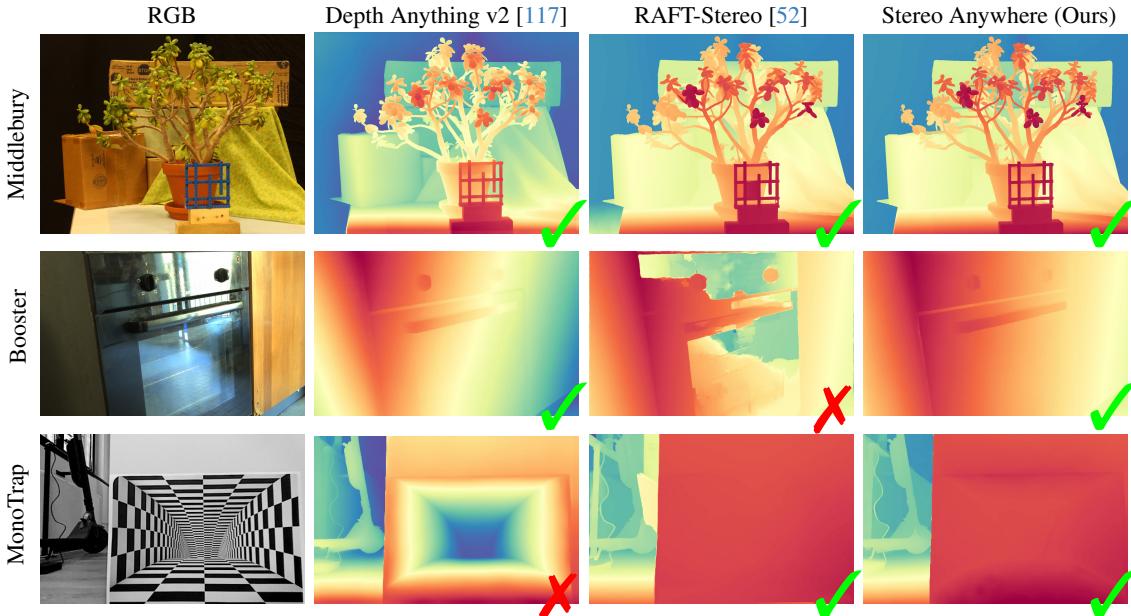


# Stereo Anywhere: Robust Zero-Shot Deep Stereo Matching Even Where Either Stereo or Mono Fail

Luca Bartolomei<sup>\*,†</sup>Fabio Tosi<sup>†</sup>Matteo Poggi<sup>\*,†</sup>Stefano Mattoccia<sup>\*,†</sup><sup>\*</sup>Advanced Research Center on Electronic System (ARCES)<sup>†</sup>Department of Computer Science and Engineering (DISI)

University of Bologna, Italy

{luca.bartolomei5, fabio.tosi5, m.poggi, stefano.mattoccia}@unibo.it

<https://stereoanywhere.github.io/>

**Figure 1. Stereo Anywhere: Combining Monocular and Stereo Strengths for Robust Depth Estimation.** Our model achieves accurate results on standard conditions (on Middlebury [83]), while effectively handling non-Lambertian surfaces where stereo networks fail (on Booster [123]) and perspective illusions that deceive monocular depth foundation models (on MonoTrap, our novel dataset).

## Abstract

We introduce *Stereo Anywhere*, a novel stereo-matching framework that combines geometric constraints with robust priors from monocular depth Vision Foundation Models (VFM). By elegantly coupling these complementary worlds through a dual-branch architecture, we seamlessly integrate stereo matching with learned contextual cues. Following this design, our framework introduces novel cost volume fusion mechanisms that effectively handle critical challenges such as textureless regions, occlusions, and non-Lambertian surfaces. Through our novel optical illusion dataset, *MonoTrap*, and extensive evaluation across multiple benchmarks, we demonstrate that our synthetic-only trained model achieves state-of-the-art results in zero-shot

generalization, significantly outperforming existing solutions while showing remarkable robustness to challenging cases such as mirrors and transparencies.

## 1. Introduction

Stereo is a fundamental task that computes depth from a synchronized, rectified image pair by finding pixel correspondences to measure their horizontal offset (*disparity*). Due to its effectiveness and minimal hardware requirements, stereo has become prevalent in numerous applications, from autonomous navigation to augmented reality.

Although in principle single-image depth estimation [3] requires an even simpler acquisition setup, its ill-posed nature leads to scale ambiguity and perspective illusion is-

sues that stereo methods inherently overcome through well-established geometric multi-view constraints.

However, despite significant advances through deep learning [44, 69], stereo models still face two main challenges: (i) limited generalization across different scenarios, and (ii) critical conditions that hinder matching or proper depth triangulation. Regarding (i), despite the initial success of synthetic datasets in enabling deep learning for stereo, their limited variety and simplified nature poorly reflect real-world complexity, and the scarcity of real training data further hinders the ability to handle heterogeneous scenarios. As for (ii), large textureless regions common in indoor environments make pixel matching highly ambiguous, while occlusions and non-Lambertian surfaces [73, 111, 123] violate the fundamental assumptions linking pixel correspondences to 3D geometry.

We argue that both challenges are rooted in the underlying limitations of stereo training data. Indeed, while data has scaled up to millions - or even billions - for several computer vision tasks, stereo datasets are still constrained in quantity and variety. This is particularly evident for non-Lambertian surfaces, which are severely underrepresented in existing datasets as their material properties prevent reliable depth measurements from active sensors (e.g. LiDAR).

In contrast, single-image depth estimation has recently witnessed a significant scale-up in data availability, reaching the order of *millions* of samples and enabling the emergence of Vision Foundation Models (VFM) [21, 40, 116, 117]. Such data abundance has influenced these models in different ways, either through direct training on large-scale depth datasets [116, 117] or indirectly by leveraging networks pre-trained on *billions* of images for diverse tasks [21, 40]. Since these models rely on contextual cues for depth estimation, they show better capability in handling textureless regions and non-Lambertian materials [72, 78, 124, 125] while being inherently immune to occlusions.

Modern graphics engines have further accelerated this progress, enabling rapid generation of high-quality synthetic data with dense depth annotations. However, although synthetic datasets featuring non-Lambertian surfaces like HyperSim [78] have proven effective for monocular depth estimation [72, 124, 125], this data abundance has not translated to stereo. Despite efforts in generating stereo pairs via novel view synthesis [23, 51, 101], available data remains insufficient for robust stereo matching.

In this paper, rather than focusing on costly real-world data collection or generating additional synthetic datasets, we propose to bridge this gap by leveraging existing VFM for single-view depth estimation. To this end, we develop a novel dual-branch deep architecture that combines stereo matching principles with monocular depth cues. Specifically, while one branch of the proposed network constructs

a cost volume from learned stereo image features, the other branch processes depth predictions from the VFM on both left and right images to build a second cost volume that incorporates depth priors to guide the disparity estimation process. These complementary signals are then iteratively combined [52], along with novel augmentation strategies applied to both cost volumes, to predict the final disparity map. Through this design, our network achieves robust performance on challenging cases like textureless regions, occlusions, and non-Lambertian surfaces, while requiring minimal synthetic stereo data. Importantly, while leveraging monocular cues, our approach preserves stereo matching geometric guarantees, effectively handling scenarios where monocular depth estimation typically fails, such as in the presence of perspective illusions. We validate this through our novel dataset of optical illusions, comprising 26 scenes with ground-truth depth maps.

We dub our framework *Stereo Anywhere*, highlighting its ability to overcome the individual limitations of stereo and monocular approaches, as depicted in Fig. 1. To summarize, our main contributions are:

- A novel deep stereo architecture leveraging monocular depth VFM to achieve strong generalization capabilities and robustness to challenging conditions.
- Novel data augmentation strategies designed to enhance the robustness of our model to textureless regions and non-Lambertian surfaces.
- A challenging dataset with optical illusion, which is particularly challenging for monocular depth with VFM.
- Extensive experiments showing Stereo Anywhere’s superior generalization and robustness to conditions critical for either stereo or monocular approaches.

## 2. Related Works

We briefly review the literature relevant to our work.

**Deep Stereo Matching.** In the last decade, stereo matching has transitioned from classical hand-crafted algorithms [82] to deep learning solutions, leading to unprecedented accuracy in depth estimation. Early deep learning efforts focused on replacing individual components of the conventional pipeline [85, 93, 102, 126, 127]. Since DispNetC [58], end-to-end architectures have evolved into 2D [50, 89, 121, 121] and 3D [4, 8, 9, 31, 41, 87, 88, 115, 128, 130] approaches, processing cost volumes through correlation layers or 3D convolutions respectively. More recent advances, thoroughly reviewed in [44, 69, 103], include recurrent architectures for stereo matching [13, 26, 37, 47, 52, 107, 112, 135] inspired by RAFT [96], Transformer-based solutions [30, 49, 56, 94, 110, 113, 133] for capturing long-range dependencies, and fully data-driven MRF models [27]. Among them, some methods specifically address temporal consistency in stereo videos [38, 39, 129, 132].

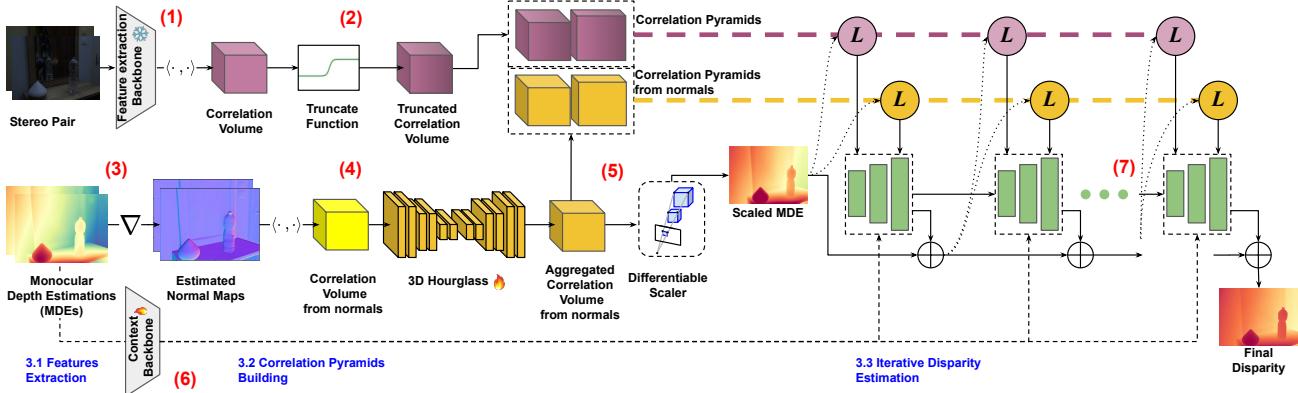


Figure 2. **Stereo Anywhere Architecture.** Given a stereo pair, (1) a pre-trained backbone is used to extract features and then build a correlation volume. Such a volume is then truncated (2) to reject matching costs computed for disparity hypotheses being *behind* non-Lambertian surfaces – glasses and mirrors. On a parallel branch, the two images are processed by a monocular VFM to obtain two depth maps (3): these are used to build a second correlation volume from retrieved normals (4). This volume is then aggregated through a 3D CNN to predict a new disparity map, used to align the original monocular depth to metric scale through a differentiable scaling module (5) for it. In parallel, the monocular depth map from left images is processed by another backbone (6) to extract context features. Finally, the two volumes and the context features from monocular depth guide the iterative disparity prediction (7).

Domain generalization remains a major challenge, with various approaches proposed including domain-invariant feature learning [16, 53, 77, 90, 131], hand-crafted matching costs [7, 14], integration of additional geometric cues [2, 63, 102], and exploitation of sparse depth measurements from active sensors [5, 46, 66]. In parallel, self-supervised approaches [24, 54] have emerged as effective alternatives to supervised learning, even using pseudo-labels from traditional algorithms [1, 97] or deploying neural radiance fields [101]. Despite the numerous attempts to improve specific aspects through the aforementioned techniques, recent architectures achieve remarkable generalization by combining their architectural advances with the increasing availability of diverse training data, while online adaptation techniques enable further improvements during deployment through self-supervised learning [42, 64, 68, 98]. However, although progress on challenges like over-smoothing [100, 114] and visually imbalanced stereo [2, 11, 55, 102], handling non-Lambertian surfaces remains particularly challenging due to limited annotated data and complex appearance, with rare works like Depth4ToM [17] specifically addressing this through semantic guidance. Among all the aforementioned approaches, there have been limited attempts to integrate stereo with monocular cues [1, 12, 109], mostly in self-supervised settings or through loose coupling between modalities.

**Monocular Depth Estimation.** Parallel to developments in stereo matching, single-image depth estimation has evolved from hand-crafted features [79] to deep learning methods [10, 20, 45, 70, 105], with self-supervised approaches [24, 25, 57, 65, 108, 134, 136] reframing the task as an image reconstruction problem. This led to multi-task

approaches incorporating flow [76, 99, 120, 137] and semantics [28, 122], alongside advances in uncertainty estimation [33, 67] and dynamic object handling [43, 60, 95]. Affine-invariant models [19, 74, 75, 118] marked a breakthrough in cross-domain generalization, pioneered by MiDaS [75] and followed by works like DPT [74] and, more recently, the Depth Anything series [116]. These approaches used different data sources, from internet photos [48, 91, 92, 118] to car sensors [22, 59] and RGB-D devices [15, 61], representing the first generation of VFMs for monocular depth estimation. Recent works have focused on metric depth estimation through camera parameter integration [29, 34, 119], diffusion models [18, 21, 36, 40, 80, 81], and temporal consistency [35, 86]. Moreover, material-aware methods [17], diffusion models [104], and large-scale synthetic datasets have enabled robust monocular depth estimation for non-Lambertian surfaces [117]. Stereo methods, however, still struggle with these surfaces due to limited real-world and synthetic annotated data, affecting generalization. We address this by integrating robust monocular VFMs into a stereo architecture.

### 3. Method Overview

Given a rectified stereo pair  $\mathbf{I}_L, \mathbf{I}_R \in \mathbb{R}^{3 \times H \times W}$ , we first obtain monocular depth estimates (MDEs)  $\mathbf{M}_L, \mathbf{M}_R \in \mathbb{R}^{1 \times H \times W}$  using a generic VFM  $\phi_M$  for monocular depth estimation. We aim to estimate a disparity map  $\mathbf{D} = \phi_S(\mathbf{I}_L, \mathbf{I}_R, \mathbf{M}_L, \mathbf{M}_R)$ , incorporating VFM priors to provide accurate results even under challenging conditions, such as texture-less areas, occlusions, and non-Lambertian surfaces. At the same time, our stereo network  $\phi_S$  is designed to avoid depth estimation errors that could arise from

relying solely on contextual cues, which can be ambiguous, like in the presence of visual illusions.

Following recent advances in iterative models [52], Stereo Anywhere comprises three main stages, as shown in Fig. 2: I) Feature Extraction, II) Correlation Pyramids Building, and III) Iterative Disparity Estimation.

### 3.1. Feature Extraction

Two distinct types of features are extracted [52]: image features and context features – (1) and (6) in Fig. 2. The image features are obtained through a feature encoder processing the stereo pair, yielding feature maps  $\mathbf{F}_L, \mathbf{F}_R \in \mathbb{R}^{D \times \frac{H}{4} \times \frac{W}{4}}$ , which are used to build a stereo correlation volume at  $\frac{1}{4}$  of the original input resolution. These encoders are initialized with pre-trained weights [52] and the image encoder is kept frozen during training. For context features, we employ a context encoder with identical architecture to the feature encoder, but processing the monocular depth estimate aligned with the reference image  $\mathbf{M}_L$  – (3) in Fig. 2 – instead of  $\mathbf{I}_L$  to capture strong geometry priors. Accordingly, during training the context encoder is optimized to extract meaningful features from these depth maps.

### 3.2. Correlation Pyramids Building

As a standard practice in stereo matching, the *cost volume* is the data structure encoding the similarity between pixels across the two images. Accordingly, our model makes use of cost volumes – in particular, of Correlation Pyramids [52] – yet in a different manner. Indeed, Stereo Anywhere builds two correlation pyramids, respectively a *stereo correlation volume* starting from  $\mathbf{I}_L, \mathbf{I}_R$  to encode image similarities, and a *monocular correlation volume* from  $\mathbf{M}_L, \mathbf{M}_R$  to encode geometric similarities – (2) and (4) in Fig. 2. Conversely to the former, the latter will not be influenced by non-Lambertian surfaces, assuming a strong  $\phi_M$ .

**Stereo Correlation Volume.** Given  $\mathbf{F}_L, \mathbf{F}_R$ , we construct a 3D correlation volume  $\mathbf{V}_S$  using dot product between feature maps:

$$(\mathbf{V}_S)_{ijk} = \sum_h (\mathbf{F}_L)_{hij} \cdot (\mathbf{F}_R)_{hik}, \quad \mathbf{V}_S \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times \frac{W}{4}} \quad (1)$$

**Monocular Correlation Volume.** Given  $\mathbf{M}_L, \mathbf{M}_R$ , we downsample them to 1/4, compute their normals  $\nabla_L, \nabla_R$ , and construct a 3D correlation volume  $\mathbf{V}_M$  using dot product between normal maps:

$$(\mathbf{V}_M)_{ijk} = \sum_h (\nabla_L)_{hij} \cdot (\nabla_R)_{hik}, \quad \mathbf{V}_M \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times \frac{W}{4}} \quad (2)$$

Given the absence of texture in  $\nabla_L$  and  $\nabla_R$ , the resulting monocular volume  $\mathbf{V}_M$  will be less informative. To alleviate this problem we segment  $\mathbf{V}_M$  using the relative depth priors from  $\mathbf{M}_L$  and  $\mathbf{M}_R$ : to do so, we generate

left and right segmentation masks  $\mathcal{M}_L \in \{0, 1\}^{\frac{H}{4} \times \frac{W}{4} \times 1}$ ,  $\mathcal{M}_R \in \{0, 1\}^{\frac{H}{4} \times 1 \times \frac{W}{4}}$ . We refer the reader to the **supplementary material** for a detailed description. Given the segmentation masks, we can generate masked volumes as:

$$(\mathbf{V}_M^n)_{ijk} = (\mathcal{M}_L^n)_{ij} \cdot (\mathcal{M}_R^n)_{ik} \cdot (\mathbf{V}_M)_{ijk} \quad (3)$$

Next, we insert a 3D Convolutional Regularization module  $\phi_A$  to aggregate  $\mathbf{V}_M^n$ , resulting in  $\mathbf{V}'_M = \phi_A(\mathbf{V}_M^1, \dots, \mathbf{V}_M^N, \mathbf{M}_L, \mathbf{M}_R)$ , with  $N = 8$ . The architecture of  $\phi_A$  follows the one in [112], with a simple permutation to match the structure of the correlation volumes. We propose an adapted version of CoEx [4] correlation volume excitation that exploits both views. The resulting feature volumes  $\mathbf{V}'_M \in \mathbb{R}^{F \times \frac{H}{4} \times \frac{W}{4} \times \frac{W}{4}}$  are fed to two different shallow 3D conv layers  $\phi_D$  and  $\phi_C$  to obtain two aggregated volumes  $\mathbf{V}_M^D = \phi_D(\mathbf{V}'_M)$  and  $\mathbf{V}_M^C = \phi_C(\mathbf{V}'_M)$  with  $\mathbf{V}_M^D, \mathbf{V}_M^C \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times \frac{W}{4}}$ .

**Differentiable Monocular Scaling.** Volume  $\mathbf{V}_M^D$  will be used not only as a monocular guide for the iterative refinement unit but also to estimate the coarse disparity maps  $\hat{\mathbf{D}}_L, \hat{\mathbf{D}}_R$ , while  $\mathbf{V}_M^C$  is used to estimate confidence maps  $\hat{\mathbf{C}}_L, \hat{\mathbf{C}}_R$ : those maps are used to scale both  $\mathbf{M}_L$  and  $\mathbf{M}_R$  – (5) in Fig. 2. To estimate left disparity from a correlation volume, we can first perform a *softargmax* on the last  $W$  dimension of  $\mathbf{V}_M^D$  to extract the correlated pixel x-coordinate, then given the relationship between left disparity and correlation  $d_L = j_L - j_R$  we obtain a coarse disparity map  $\hat{\mathbf{D}}_L$ :

$$(\hat{\mathbf{D}}_L)_{ij} = j - \text{softargmax}_L(\mathbf{V}_M^D)_{ij} \quad (4)$$

Similarly, we can estimate  $\hat{\mathbf{D}}_R$  from  $\mathbf{V}_M^D$ . We refer the reader to the supplementary for details. We also aim to estimate a pair of confidence maps  $\hat{\mathbf{C}}_L, \hat{\mathbf{C}}_R \in [0, 1]^{H \times W}$  to classify outliers and perform a robust scaling. Inspired by information entropy, we estimate the *chaos* inside correlation curves: clear monomodal-like cost curve – *i.e.*, the ones with low entropy – are reliable – while *chaotic* curves – *i.e.*, the ones with high entropy – are uncertain. To estimate the left confidence map, we perform a *softmax* operation on the last  $W$  dimension of  $\mathbf{V}_M^C$ , then  $\hat{\mathbf{C}}_L$  is obtained as follows:

$$(\hat{\mathbf{C}}_L)_{ij} = 1 + \frac{\sum_d^{\frac{W}{4}} \frac{e^{(\mathbf{V}_M^C)_{ijd}}}{\sum_f^{\frac{W}{4}} e^{(\mathbf{V}_M^C)_{ifd}}} \cdot \log_2 \left( \frac{e^{(\mathbf{V}_M^C)_{ijd}}}{\sum_f^{\frac{W}{4}} e^{(\mathbf{V}_M^C)_{ifd}}} \right)}{\log_2(\frac{W}{4})} \quad (5)$$

In the same way, we estimate  $\hat{\mathbf{C}}_R$ . To further reduce outliers, we mask out from  $\hat{\mathbf{C}}_L$  and  $\hat{\mathbf{C}}_R$  occluded pixels using a *SoftLRC* operator – see the **supplementary material** for details. Finally, we can estimate the scale  $\hat{s}$  and shift  $\hat{t}$  using a differentiable weighted least-square approach:

$$\min_{\hat{s}, \hat{t}} \sum_{i,j}^{L,R} \left\| \sqrt{\hat{\mathbf{C}}} \odot [(\hat{s}\mathbf{M} + \hat{t}) - \hat{\mathbf{D}}] \right\|_F \quad (6)$$

where  $\|\cdot\|_F$  denotes the Frobenius norm. Using the scaling coefficients, we obtain two disparity maps  $\hat{\mathbf{M}}_L, \hat{\mathbf{M}}_R$ :

$$\hat{\mathbf{M}}_L = \hat{s}\mathbf{M}_L + \hat{t}, \quad \hat{\mathbf{M}}_R = \hat{s}\mathbf{M}_R + \hat{t} \quad (7)$$

It is crucial to optimize both left and right scaling jointly to obtain consistency between  $\hat{\mathbf{M}}_L$  and  $\hat{\mathbf{M}}_R$ .

**Volume Augmentations.** Unfortunately, Stereo Anywhere cannot properly learn when to choose stereo or mono information from [58] alone. Hence, we propose three volume augmentations and a monocular augmentation to overcome this issue: 1) *Volume Rolling*: we randomly apply a rolling operation to the last  $W$  dimension of  $\mathbf{V}^D_M$  or  $\mathbf{V}_S$ ; 2) *Volume Noising*: we apply random noise sampled from the interval  $[0, 1]$  using a uniform distribution; 3) *Volume Zeroing*: we apply a Gaussian-like curve with the peak where disparity equals zero. Furthermore, we randomly substitute the monocular prediction with the ground truth normalized between  $[0, 1]$  as an additional augmentation. We apply only one volume augmentation to  $\mathbf{V}^D_M$  or  $\mathbf{V}_S$  and only for a section of the volume, randomly selecting an  $\mathcal{M}_L^n$  mask.

**Volume Truncation.** To further help Stereo Anywhere to handle mirror surfaces, we introduce a hand-crafted volume truncation operation on  $\mathbf{V}_S$ . Firstly, we extract left confidence  $\mathbf{C}_M = \text{softLRC}_L(\hat{\mathbf{M}}_L, \hat{\mathbf{M}}_R)$  to classify reliable monocular predictions. Then, we create a truncate mask  $\mathbf{T} \in [0, 1]^{\frac{H}{4} \times \frac{W}{4}}$  using the following logic condition:  $(\mathbf{T})_{ij} = \left[ \left( (\hat{\mathbf{M}}_L)_{ij} > (\hat{\mathbf{D}}_L)_{ij} \right) \wedge (\mathbf{C}_M)_{ij} \right] \vee \left[ (\mathbf{C}_M)_{ij} \wedge \neg(\hat{\mathbf{C}}_L)_{ij} \right]$ . We implement this logic using fuzzy operators (more details in the **supplementary material**). The rationale is that stereo predicts farther depths on mirror surfaces: the mirror is perceived as a window on a new environment, specular to the real one. Finally, for values of  $\mathbf{T} > T_m = 0.98$ , we truncate  $\mathbf{V}_S$  using a sigmoid curve centered at the correlation value predicted by  $\hat{\mathbf{M}}_L$  – i.e., the real disparity of mirror surfaces – preserving only the stereo correlation curve that does not “pierce” the mirror region.

### 3.3. Iterative Disparity Estimation

We aim to estimate a series of refined disparity maps  $\{\mathbf{D}^1 = \hat{\mathbf{M}}_L, \mathbf{D}^2, \dots, \mathbf{D}^l, \dots\}$  exploiting the guidance from both stereo and mono branches. Starting from the Multi-GRU update operator by [52], we introduce a second lookup operator that extracts correlation features  $\mathbf{G}_M$  from the additional volume  $\mathbf{V}_M^D$  – (7) in Fig. 2. The two sets of correlation features from  $\mathbf{G}_S$  and  $\mathbf{G}_M$  are processed by the same two-layer encoder and concatenated with features derived from the current disparity estimation  $\mathbf{D}^l$ . This concatenation is further processed by a 2D conv layer, and then by the ConvGRU operator. We inherit the convex upsampling module [52] to upsample final disparity to full resolution.

### 3.4. Training Supervision

We supervise the iterative module using the well-known L1 loss with exponentially increasing weights [52], then  $\hat{\mathbf{D}}_L, \hat{\mathbf{D}}_R, \hat{\mathbf{M}}_L$  and  $\hat{\mathbf{M}}_R$  using the L1 loss, finally  $\hat{\mathbf{C}}_L$  and  $\hat{\mathbf{C}}_R$  using the Binary Cross Entropy loss. We invite the reader to read the **supplementary material** for additional details.

## 4. The MonoTrap Dataset

Monocular depth estimation is known for possibly failing in the presence of perspective illusions. The reader may wonder how Stereo Anywhere would behave in such cases: would it blindly trust the monocular VFM or rely on the stereo geometric principles to maintain robustness?

To answer these questions, we introduce MonoTrap, a novel stereo dataset specifically designed to challenge monocular depth estimation. Our dataset comprises 26 scenes featuring perspective illusions, captured with a calibrated stereo setup and annotated with ground-truth depth from an Intel Realsense L515 LiDAR. The scenes contain carefully designed planar patterns that create visual illusions, such as apparent holes in walls or floors and simulated transparent surfaces that reveal content behind them. Figure 3 shows examples from our dataset that illustrate how these visual illusions easily fool monocular methods.

## 5. Experiments

We describe our implementation details, datasets, and evaluation protocols, followed by experiments. We also refer the reader to the **supplementary material** for more results.

### 5.1. Implementation and Experimental Settings

We implement Stereo Anywhere using PyTorch, starting from RAFT-Stereo codebase [52]. We use Depth Anything v2 [117] as the VFM fueling our model, using the *Large* weights provided by the authors, trained on ground-truth labels from the HyperSim synthetic dataset [78] only.

Starting from the Sceneflow RAFT-Stereo checkpoint, we train Stereo Anywhere on a single A100 GPU for 3 epochs, with learning rate 1e-4 and AdamW optimizer, on batches of 2 images. We extract random crops of size  $320 \times 640$  from images and apply standard color and spatial augmentations [52]. The VMF is used only to source monocular depth maps, remaining frozen during training.

The number of iterations for GRUs is fixed to 12 during training and increased to 32 at inference time.

### 5.2. Evaluation Datasets & Protocol

**Datasets.** We utilize SceneFlow [58] as our sole training dataset, comprising about 39k synthetic stereo pairs with dense ground-truth disparities. For evaluation, we employ several benchmarks: Middlebury 2014 [83] and its 2021

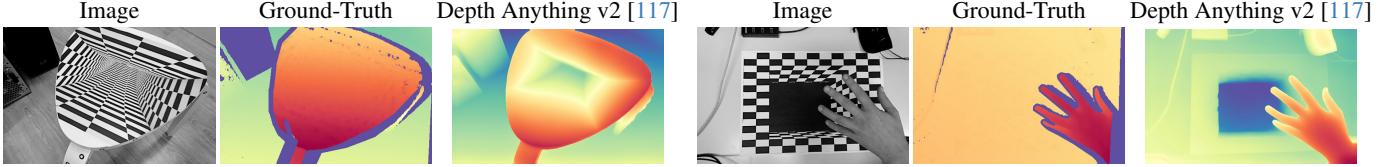


Figure 3. **Samples from MonoTrap Dataset.** We report two scenes featured in our dataset, showing the left image, the ground-truth depth, and the predictions by Depth Anything v2 [117], highlighting how it fails in the presence of visual illusions.

Experiment	Booster (Q)					Middlebury 2014 (H)				
	> 2	> 4	bad	> 6	> 8	Avg. (px)	All	bad > 2 Noc	Occ	Avg. (px)
(A) Baseline [52]	17.86	13.09	10.76	9.24	3.58	11.15	8.06	29.06	1.55	
(B) (A) + Monocular Context w/o re-train	15.85	10.98	8.89	7.69	3.05	14.96	11.70	34.38	2.82	
(C) (A) + Monocular Context w/ re-train	14.94	10.40	8.61	7.63	3.03	9.62	6.98	25.39	1.13	
(D) (C) + Normals Correlation Volume / Scaled Depth	11.33	6.88	5.32	4.59	1.87	7.67	5.24	21.51	<b>0.96</b>	
(E) (D) + Volume augmentation / truncation	<b>9.96</b>	<b>5.81</b>	<b>4.48</b>	<b>3.79</b>	<b>1.36</b>	<b>7.07</b>	<b>4.76</b>	<b>20.77</b>	0.97	

Table 1. **Ablation Studies.** We measure the impact of different design strategies. Networks trained on SceneFlow [58].

extension [62] provide high-resolution indoor scenes with semi-dense labels (15 and 24 stereo pairs), KITTI 2012 [22] and 2015 [59] feature outdoor driving scenarios ( $\sim 200$  pairs each at  $1280 \times 384$  with sparse LiDAR ground truth), and ETH3D [84] contributes 27 low-resolution indoor/outdoor scenes. For non-Lambertian surfaces, we primarily use Booster [123], containing 228 high-resolution (12 Mpx) indoor pairs with its 191-pair online benchmark, and LayeredFlow [111], featuring 400 pairs with transparent objects and sparse ground truth ( $\sim 50$  points per pair). Additionally, we include our newly proposed MonoTrap dataset focusing on optical illusions. For zero-shot evaluation, we test on KITTI 2015, Middlebury v3 at half (H) resolution, Middlebury 2021, and ETH3D, while non-Lambertian zero-shot testing relies on Booster at quarter (Q) resolution and LayeredFlow at eight (E) resolution.

**Evaluation Metrics.** We evaluate our method using two standard metrics: the average pixel error (Avg.), which computes the absolute difference between predicted and ground truth disparities averaged over all pixels, and the bad- $\tau$  error, which measures the percentage of pixels with a disparity error greater than  $\tau$  pixels – for the latter, we compute it considering all pixels or either non-occluded or occluded pixels, referred to as *All*, *Noc* or *Occ* respectively.

We evaluate on MonoTrap through standard monocular depth metrics [24] - Absolute relative error (AbsRel), RMSE, and  $\delta < 1.05$  score.

### 5.3. Ablation Study

We start our analysis by evaluating how individual components of our model contribute to the overall accuracy. All model variants are trained solely on the synthetic SceneFlow dataset and tested on Booster and Middlebury 2014, allowing us to examine their effectiveness on non-Lambertian surfaces and general scenes.

Table 1 summarizes our findings. In (A), we report the

performance of our baseline model, upon which we build Stereo Anywhere– i.e., RAFT-stereo [52]. On the one hand, by adding monocular context from an off-the-shelf monocular depth network to the pre-trained context backbone (B), we observe improved performance on non-Lambertian surfaces, though at the expense of a general drop in accuracy on Middlebury. On the other hand, by re-training the context backbone to process depth maps obtained from the monocular network on SceneFlow (C), we can appreciate a consistent improvement in both datasets. Introducing the normals correlation volume with subsequent differentiable depth scaling (D) significantly enhances the accuracy on non-Lambertian surfaces, also showing improvements on indoor scenes. Finally, cost volume augmentations and truncation (E) demonstrate positive effects on transparent surfaces and mirrors present in the Booster dataset by further reducing the bad-2 metric by approximately 1.5% and Avg. by 0.5 pixels, with minimal influence on Middlebury.

According to these results, from now on, we will adopt (E) as the default setting for Stereo Anywhere.

### 5.4. Zero-Shot Generalization

We now compare our Stereo Anywhere model against state-of-the-art deep stereo networks, assessing zero-shot generalization capability when transferred from synthetic to real images. Purposely, we follow a well-established benchmark in the literature [52, 101], evaluating on real datasets models pre-trained exclusively on SceneFlow [58].

Table 2 compares Stereo Anywhere with off-the-shelf stereo networks using authors’ provided weights. Considering All, Noc, and Avg. metrics, we can notice how Stereo Anywhere achieves consistently better results across most datasets, achieving almost 3% lower bad-2 All on Middlebury 2014 versus the second-best method DLNR [135], and breaking the 4% barrier on KITTI’s bad-3 All metric.

The Occ metric further demonstrates how Stereo Any-

Model	Middlebury 2014 (H)						Middlebury 2021						ETH3D			KITTI 2012			KITTI 2015					
	bad > 2			Avg.			bad > 2			Avg.			bad > 1			Avg.			bad > 3			Avg.		
	All	Noc	Occ	(px)	All	Noc	Occ	(px)	All	Noc	Occ	(px)	All	Noc	Occ	(px)	All	Noc	Occ	(px)	All	Noc	Occ	(px)
RAFT-Stereo [52]	11.15	8.06	29.06	1.55	12.05	9.38	37.89	1.81	2.59	2.24	8.78	0.25	4.80	3.70	28.54	0.89	5.44	4.69	13.52	1.16				
PSMNet [8]	18.79	13.80	53.22	4.63	23.67	20.61	53.75	5.70	19.75	18.62	42.05	0.94	6.73	5.28	45.48	1.22	6.78	5.84	24.22	1.38				
GMStereo [113]	15.63	10.98	46.04	1.87	25.43	22.43	54.70	2.86	6.22	5.58	19.97	0.42	5.68	4.34	38.12	1.10	5.72	4.92	16.74	1.21				
ELFNet [56]	24.48	16.94	77.06	8.61	27.08	21.77	85.56	11.01	25.61	24.50	46.06	5.65	10.52	8.13	87.24	2.30	9.61	7.67	84.71	2.16				
PCVNet [128]	18.50	14.73	42.16	3.45	18.40	15.06	51.50	3.84	7.81	7.19	20.07	0.69	5.14	3.95	32.07	0.95	5.49	4.74	15.14	1.34				
DLNR [135]	9.46	6.20	28.75	1.45	8.44	5.88	32.71	1.24	23.12	22.94	26.93	9.89	9.45	8.30	36.05	1.59	15.74	14.87	33.65	2.83				
Selective-RAFT [107]	12.05	9.46	27.42	2.35	15.69	13.86	36.32	5.92	4.36	3.81	10.23	0.34	5.71	4.63	29.87	1.08	6.50	5.69	17.85	1.27				
Selective-IGEV [107]	9.98	7.09	27.62	1.60	8.89	6.34	32.88	1.60	6.42	5.71	18.71	1.73	6.22	4.91	34.08	1.09	5.87	5.15	14.42	1.42				
IGEV-Stereo [112]	15.07	11.81	35.78	3.20	20.43	18.14	45.37	8.16	43.05	42.42	57.19	1.04	7.62	5.90	56.13	1.50	7.81	6.68	42.29	1.56				
NMRF [27]	14.08	10.87	34.62	2.91	23.36	21.69	42.51	8.57	4.34	3.66	17.15	0.42	4.62	3.52	29.98	0.92	5.24	4.55	11.72	1.16				
Stereo Anywhere (ours)	<b>7.07</b>	<b>4.76</b>	<b>20.77</b>	<b>0.97</b>	<b>8.38</b>	<b>5.86</b>	<b>32.87</b>	<b>1.10</b>	<b>2.39</b>	<b>2.16</b>	<b>5.82</b>	<b>0.28</b>	<b>3.94</b>	<b>3.03</b>	<b>21.02</b>	<b>0.85</b>	<b>3.98</b>	<b>3.29</b>	<b>10.55</b>	<b>0.97</b>				

Table 2. **Zero-shot Generalization.** Comparison with state-of-the-art deep stereo models. Networks trained on SceneFlow [58].

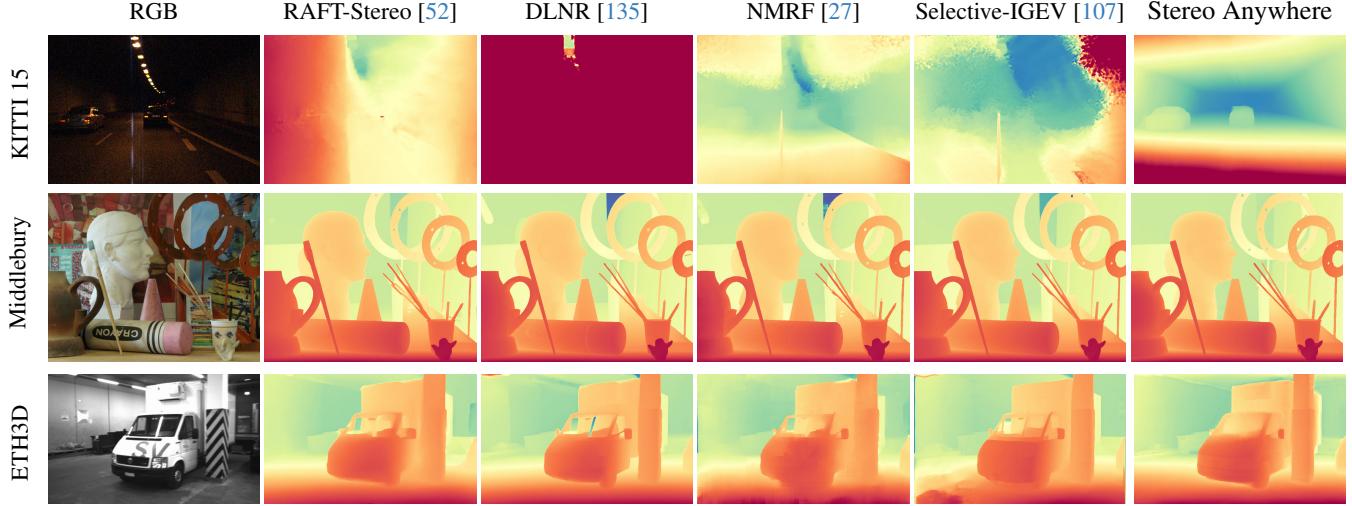


Figure 4. **Qualitative Results – Zero-Shot Generalization.** Predictions by state-of-the-art models and Stereo Anywhere.

where consistently outperforms other stereo models on any dataset, with substantial margins over the second-best – i.e., approximately 7% on Middlebury 2014 and KITTI 2012, and 3% on ETH3D. This confirms that leveraging priors from VFM for monocular depth estimation effectively improve the stereo matching estimation accuracy in challenging conditions where stereo matching is ill-posed, such as at occluded regions.

Figure 4 shows predictions on KITTI 2015, Middlebury 2014, and ETH3D samples. In particular, the first row shows an extremely challenging case for SceneFlow-trained models, where Stereo Anywhere achieves accurate disparity maps thanks to VFM priors.

### 5.5. Zero-Shot Non-Lambertian Generalization

We now assess the generalization capabilities of Stereo Anywhere and existing stereo models when dealing with non-Lambertian materials, such as transparent surfaces or mirrors. To this end, we conduct a zero-shot generalization evaluation experiment on the Booster [71] and Layered-Flow [111] datasets, once again using models pre-trained on SceneFlow [58] – with weights provided by the authors.

Table 3 shows the outcome of this evaluation. This time, we can perceive even more clearly how Stereo Anywhere is

the absolute winner, demonstrating unprecedented robustness in the presence of non-Lambertian surfaces despite being trained only on synthetic stereo data, not even featuring such objects. These results further validate how leveraging strong priors from existing VFM for monocular depth estimation can play a game-changing role in stereo matching as well, especially when lacking training data explicitly targeting critical conditions such as non-Lambertian surfaces.

Figure 5 shows examples from Booster and Layered-Flow, where Stereo Anywhere is the only stereo model correctly perceiving the mirror and transparent railing.

### 5.6. MonoTrap Benchmark

We conclude our evaluation by running experiments on our newly collected MonoTrap dataset to prove the robustness of Stereo Anywhere in the presence of critical conditions harming the accuracy of monocular depth predictors.

Table 4 collects the results achieved by state-of-the-art monocular depth estimation models, the baseline stereo model over which we built our framework (RAFT-Stereo) and Stereo Anywhere. Regarding the former models, as they predict affine-invariant depth maps, following the literature [75] we use least square errors to align them to the ground-truth. As these models are fooled by the visual il-

Model	Booster (Q)				Avg. (px)	LayeredFlow (E)			Avg. (px)
	> 2	Error Rate (%) > 4	> 6	> 8		> 1	Error Rate (%) > 3	> 5	
RAFT-Stereo [52]	17.84	13.06	10.76	9.24	3.59	89.21	79.02	71.61	19.27
PSMNet [8]	34.47	24.83	20.46	17.77	7.26	91.85	79.84	70.04	21.18
GMStereo [113]	32.44	22.52	17.96	15.02	5.29	92.95	83.68	74.76	20.91
ELFNNet [56]	45.52	35.79	30.72	27.33	14.04	93.08	82.24	70.41	20.19
PCVNet [128]	31.08	22.90	18.80	16.15	6.16	91.64	80.75	74.34	20.85
DLNer [135]	18.56	14.55	12.61	11.22	3.97	89.90	79.46	72.72	18.97
Selective-RAFT [107]	20.01	15.08	12.52	10.88	4.12	92.69	86.32	78.82	20.18
Selective-IGEV [107]	18.52	14.24	12.14	10.77	4.38	91.31	81.72	74.74	19.65
IGEV-Stereo [112]	23.38	14.45	12.61	11.41	4.91	92.54	81.42	74.74	20.88
NMRF [27]	27.08	19.06	15.43	13.21	5.02	89.08	79.13	70.51	20.17
<b>Stereo Anywhere (ours)</b>	<b>9.96</b>	<b>5.81</b>	<b>4.48</b>	<b>3.79</b>	<b>1.36</b>	<b>80.83</b>	<b>58.21</b>	<b>46.48</b>	<b>12.14</b>

Table 3. **Zero-shot Non-Lambertian Generalization.** Comparison with state-of-the-art models. Networks trained on SceneFlow [58].

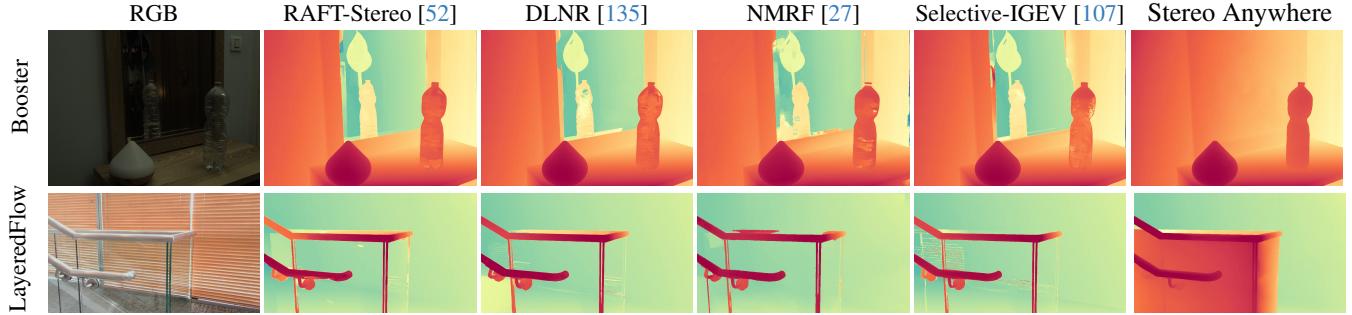


Figure 5. **Qualitative results – Zero-Shot non-Lambertian Generalization.** Predictions by state-of-the-art models and Stereo Anywhere.

Model	MonoTrap		
	AbsRel (%)↓	RMSE (m)↓	$\sigma < 1.05$ (%)↑
Depth Anything v2 [117]	35.00	0.34	27.62
Depth Anything v2 [117] †	21.81	0.28	29.31
DepthPro [6]	29.96	0.30	27.81
DepthPro [6] †	17.23	0.25	37.85
RAFT-Stereo [52]	4.62	0.12	75.15
<b>Stereo Anywhere</b>	<b>4.59</b>	<b>0.11</b>	<b>77.26</b>

Table 4. **MonoTrap Benchmark.** Comparison with state-of-the-art monocular depth estimation models and RAFT-Stereo. Both RAFT-Stereo and Stereo Anywhere are trained on SceneFlow [58]. † refers to robust scaling through RANSAC.

lusions, this scaling procedure is likely to yield sub-optimal scale and shift parameters. Therefore, we alternatively align to ground-truth depth through a more robust RANSAC fitting – denoted with † in the table.

On the one hand, by comparing monocular and stereo methods, we notice how the failures of the former negatively impact their evaluation metrics. Once again, we remark that a direct comparison across the two families of methods is not the main goal of this experiment. On the other hand, we focus on the comparison between RAFT-Stereo and Stereo Anywhere, with our model performing slightly better than its baseline. This fact proves that despite its strong reliance on the priors retrieved from VFM for monocular depth estimation, Stereo Anywhere can properly ignore such priors when unreliable.

Figure 6 shows three samples where Depth Anything v2

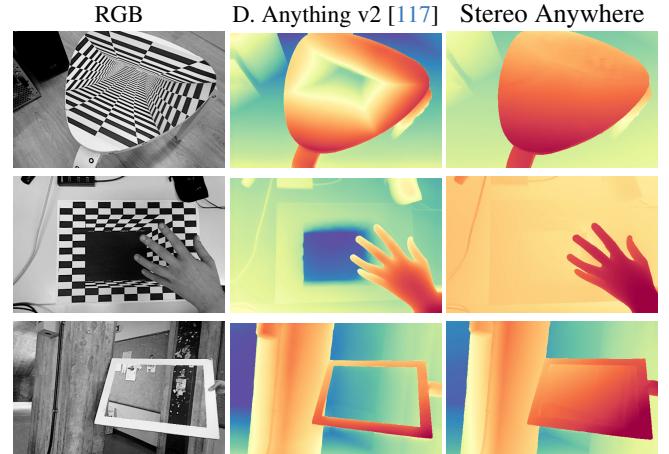


Figure 6. **Qualitative results – MonoTrap.** Stereo Anywhere is not fooled by erroneous predictions by its monocular engine [117].

fails while Stereo Anywhere does not.

## 6. Conclusion

In this paper, we introduced Stereo Anywhere, a novel stereo matching framework that leverages monocular depth VFM to overcome traditional stereo matching limitations. Combining stereo geometric constraints with monocular priors, our approach demonstrates superior zero-shot generalization and robustness to challenging conditions like textureless regions, occlusions, and non-Lambertian sur-

faces. Furthermore, through our novel MonoTrap dataset, we showed that Stereo Anywhere effectively combines the best of both worlds - maintaining stereo matching’s geometric accuracy where monocular methods fail, while leveraging monocular priors to handle challenging stereo scenarios. Extensive comparisons against state-of-the-art networks in zero-shot settings validate these findings.

## References

- [1] Filippo Aleotti, Fabio Tosi, Li Zhang, Matteo Poggi, and Stefano Mattoccia. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 614–632. Springer, 2020. [3](#)
- [2] Filippo Aleotti, Fabio Tosi, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Neural disparity refinement for arbitrary resolution stereo. In *2021 International Conference on 3D Vision (3DV)*, pages 207–217. IEEE, 2021. [3](#)
- [3] Vasileios Arampatzakis, George Pavlidis, Nikolaos Mitanoudis, and Nikos Papamarkos. Monocular depth estimation: A thorough review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. [1](#)
- [4] Antyanta Bangunharcana, Jae Won Cho, Seokju Lee, In So Kweon, Kyung-Soo Kim, and Soohyun Kim. Correlate-and-excite: Real-time stereo matching via guided cost volume excitation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. [2, 4, 15](#)
- [5] Luca Bartolomei, Matteo Poggi, Fabio Tosi, Andrea Conti, and Stefano Mattoccia. Active stereo without pattern projector. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18470–18482, 2023. [3](#)
- [6] 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*, 2024. [8, 18, 19, 31](#)
- [7] Changjiang Cai, Matteo Poggi, Stefano Mattoccia, and Philippas Mordohai. Matching-space stereo networks for cross-domain generalization. In *2020 International Conference on 3D Vision (3DV)*, pages 364–373, 2020. [3](#)
- [8] 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, 8](#)
- [9] Liyan Chen, Weihan Wang, and Philippas Mordohai. Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17235–17244, 2023. [2](#)
- [10] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 730–738, Red Hook, NY, USA, 2016. Curran Associates Inc. [3](#)
- [11] Xihao Chen, Zhiwei Xiong, Zhen Cheng, Jiayong Peng, Yueyi Zhang, and Zheng-Jun Zha. Degradation-agnostic correspondence from resolution-asymmetric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12962–12971, 2022. [3](#)
- [12] Zhi Chen, Xiaoqing Ye, Wei Yang, Zhenbo Xu, Xiao Tan, Zhikang Zou, Errui Ding, Xinming Zhang, and Liusheng Huang. Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15529–15538, 2021. [3](#)
- [13] Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bing-shu Wang, Yongbin Qin, and Jia Wu. Mocha-stereo: Motif channel attention network for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. [2](#)
- [14] Kelvin Cheng, Tianfu Wu, and Christopher Healey. Revisiting non-parametric matching cost volumes for robust and generalizable stereo matching. *Advances in Neural Information Processing Systems*, 35:16305–16318, 2022. [3](#)
- [15] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. Diml/cvl rgb-d dataset: 2m rgb-d images of natural indoor and outdoor scenes. *arXiv preprint arXiv:2110.11590*, 2021. [3](#)
- [16] WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar, and David Suter. Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13022–13032, 2022. [3](#)
- [17] Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Learning depth estimation for transparent and mirror surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9244–9255, 2023. [3](#)
- [18] Yiqun Duan, Xianda Guo, and Zheng Zhu. DiffusionDepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023. [3](#)
- [19] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10786–10796, 2021. [3](#)
- [20] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, page 2366–2374, Cambridge, MA, USA, 2014. MIT Press. [3](#)
- [21] 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, 3](#)
- [22] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. [3, 6](#)

- [23] Magnus Kaufmann Gjerde, Filip Slezák, Joakim Bruslund Haurum, and Thomas B Moeslund. From nerf to 3dgs: A leap in stereo dataset quality? In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024. 2
- [24] Clément Godard, Oisin 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, 6
- [25] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019. 3
- [26] Rui Gong, Weide Liu, Zaiwang Gu, Xulei Yang, and Jun Cheng. Learning intra-view and cross-view geometric knowledge for stereo matching. *arXiv preprint arXiv:2402.19270*, 2024. 2
- [27] Tongfan Guan, Chen Wang, and Yun-Hui Liu. Neural markov random field for stereo matching, 2024. 2, 7, 8, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30
- [28] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*, 2020. 3
- [29] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareş Ambrus, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *ICCV*, 2023. 3
- [30] Weiyu Guo, Zhaoshuo Li, Yongkui Yang, Zheng Wang, Russell H Taylor, Mathias Unterath, Alan Yuille, and Yingwei Li. Context-enhanced stereo transformer. In *European Conference on Computer Vision*, pages 263–279. Springer, 2022. 2
- [31] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, 2019. 2
- [32] 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. 18, 19
- [33] Julia Hornauer and Vasileios Belagiannis. Gradient-based uncertainty for monocular depth estimation. In *European Conference on Computer Vision*, pages 613–630. Springer, 2022. 3
- [34] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 3
- [35] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 3
- [36] 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 *ICCV*, 2023. 3
- [37] Junpeng Jing, Jiankun Li, Pengfei Xiong, Jiangyu Liu, Shuaicheng Liu, Yichen Guo, Xin Deng, Mai Xu, Lai Jiang, and Leonid Sigal. Uncertainty guided adaptive warping for robust and efficient stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3318–3327, 2023. 2
- [38] Junpeng Jing, Ye Mao, and Krystian Mikolajczyk. Match-stereo-videos: Bidirectional alignment for consistent dynamic stereo matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [39] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13229–13239, 2023. 2
- [40] 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 (CVPR)*, 2024. 2, 3
- [41] 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
- [42] Kwonyoung Kim, Jungin Park, Jiyoung Lee, Dongbo Min, and Kwanghoon Sohn. Pointfix: Learning to fix domain bias for robust online stereo adaptation. In *European Conference on Computer Vision*, pages 568–585. Springer, 2022. 3
- [43] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 582–600. Springer, 2020. 3
- [44] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):1738–1764, 2020. 2
- [45] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 3
- [46] Ang Li, Anning Hu, Wei Xi, Wenxian Yu, and Danding Zou. Stereo-lidar depth estimation with deformable propagation and learned disparity-depth conversion. *arXiv preprint arXiv:2404.07545*, 2024. 3
- [47] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022. 2

- [48] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. 3
- [49] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Uebelath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6197–6206, 2021. 2
- [50] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2811–2820, 2018. 2
- [51] Han Ling, Yinghui Sun, Quansen Sun, Ivor Tsang, and Yuhui Zheng. Self-assessed generation: Trustworthy label generation for optical flow and stereo matching in real-world. *arXiv preprint arXiv:2410.10453*, 2024. 2
- [52] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, 2021. 1, 2, 4, 5, 6, 7, 8, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
- [53] Biyang Liu, Huimin Yu, and Guodong Qi. Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13012–13021, 2022. 3
- [54] Pengpeng Liu, Irwin King, Michael R. Lyu, and Jia Xu. Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [55] Yicun Liu, Jimmy Ren, Jiawei Zhang, Jianbo Liu, and Mude Lin. Visually imbalanced stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2029–2038, 2020. 3
- [56] Jieming Lou, Weide Liu, Zhuo Chen, Fayao Liu, and Jun Cheng. Elfnet: Evidential local-global fusion for stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17784–17793, 2023. 2, 7, 8
- [57] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5667–5675, 2018. 3
- [58] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 2, 5, 6, 7, 8, 16, 18
- [59] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 3, 6
- [60] Jaeho Moon, Juan Luis Gonzalez Bello, Byeongjun Kwon, and Munchurl Kim. From-ground-to-objects: Coarse-to-fine self-supervised monocular depth estimation of dynamic objects with ground contact prior. *arXiv preprint arXiv:2312.10118*, 2023. 3
- [61] Pushmeet Kohli, Nathan Silberman, Derek Hoiem, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 3
- [62] Guanghan Pan, Tiansheng Sun, Toby Weed, and Daniel Scharstein. 2021 Mobile stereo datasets with ground truth. <https://vision.middlebury.edu/stereo/data/scenes2021/>, 2021. 6
- [63] Andrea Pilzer, Yuxin Hou, Niki Loppi, Arno Solin, and Juho Kannala. Expansion of visual hints for improved generalization in stereo matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5840–5849, 2023. 3
- [64] Matteo Poggi and Fabio Tosi. Federated online adaptation for deep stereo. In *CVPR*, 2024. 3
- [65] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *2018 International conference on 3d vision (3DV)*, pages 324–333. IEEE, 2018. 3
- [66] Matteo Poggi, Davide Pallotti, Fabio Tosi, and Stefano Mattoccia. Guided stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 979–988, 2019. 3
- [67] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3237, 2020. 3
- [68] Matteo Poggi, Alessio Tonioni, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Continual adaptation for deep stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4713–4729, 2021. 3
- [69] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippo Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5314–5334, 2021. 2
- [70] Michael Ramamonjisoa, Yuming Du, and Vincent Lepetit. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In *CVPR*, 2020. 3
- [71] Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: The booster dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21168–21178, 2022. 7
- [72] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, Samuele Salti, Stefano Mattoccia, Jun Shi, Dafeng Zhang, Yong A, Yixiang Jin, Dingzhe Li, Chao Li, Zhiwen Liu, Qi Zhang, Xinxing Wang, and Shi Yin. Ntire 2023 challenge on hr

- depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023. CVPRW. 2
- [73] 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. 2
- [74] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. 3
- [75] 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. 3, 7
- [76] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12240–12249, 2019. 3
- [77] Zhibo Rao, Bangshu Xiong, Mingyi He, Yuchao Dai, Renjie He, Zhelun Shen, and Xing Li. Masked representation learning for domain generalized stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5435–5444, 2023. 3
- [78] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atul Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 2, 5
- [79] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Depth perception from a single still image. In *Proc. AAAI*, 2008. 3
- [80] 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. *arXiv preprint arXiv:2306.01923*, 2023. 3
- [81] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 3
- [82] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002. 2
- [83] 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, GCP 2014, Münster, Germany, September 2–5, 2014, Proceedings 36*, pages 31–42. Springer, 2014. 1, 5
- [84] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 6
- [85] Akihito Seki and Marc Pollefeys. Sgm-nets: Semi-global matching with neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 231–240, 2017. 2
- [86] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. *arXiv preprint arXiv:2406.01493*, 2024. 3
- [87] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2021. 2
- [88] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *European Conference on Computer Vision*, pages 280–297. Springer, 2022. 2
- [89] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 20–35. Springer, 2019. 2
- [90] Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Zhe Wang, and Jianping Shi. Adastereo: A simple and efficient approach for adaptive stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10328–10337, 2021. 3
- [91] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back & relax: Learning to reconstruct the world by watching slowtv. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15768–15779, 2023. 3
- [92] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back & relax++: Scaling beyond ground-truth depth with slowtv & cribstv. *arXiv preprint arXiv:2403.01569*, 2024. 3
- [93] Aristotle Spyropoulos, Nikos Komodakis, and Philippos Mordohai. Learning to detect ground control points for improving the accuracy of stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1621–1628, 2014. 2
- [94] Qing Su and Shihao Ji. Chittransformer: Towards reliable stereo from cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1939–1949, 2022. 2
- [95] Yihong Sun and Bharath Hariharan. Dynamo-depth: Fixing unsupervised depth estimation for dynamical scenes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3
- [96] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, Au-*

- gust 23–28, 2020, Proceedings, Part II 16, pages 402–419. Springer, 2020. 2
- [97] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised adaptation for deep stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1605–1613, 2017. 3
- [98] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 195–204, 2019. 3
- [99] Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Distilled semantics for comprehensive scene understanding from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4654–4665, 2020. 3
- [100] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3
- [101] Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. Nerf-supervised deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 855–866, 2023. 2, 3, 6
- [102] Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Neural disparity refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 3
- [103] Fabio Tosi, Luca Bartolomei, and Matteo Poggi. A survey on deep stereo matching in the twenties. *arXiv preprint arXiv:2407.07816*, 2024. Extended version of CVPR 2024 Tutorial "Deep Stereo Matching in the Twenties" (<https://sites.google.com/view/stereo-twenties>). 2
- [104] Fabio Tosi, Pierluigi Zama Ramirez, and Matteo Poggi. Diffusion models for monocular depth estimation: Overcoming challenging conditions. In *European Conference on Computer Vision (ECCV)*, 2024. 3
- [105] Lijun Wang, Jianming Zhang, Yifan Wang, Huchuan Lu, and Xiang Ruan. CLIFFNet for monocular depth estimation with hierarchical embedding loss. In *ECCV*. Springer, 2020. 3
- [106] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision, 2024. 18, 19
- [107] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 7, 8, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
- [108] Jamie Watson, Michael Firman, Gabriel J. Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *ICCV*, 2019. 3
- [109] Jamie Watson, Oisin Mac Aodha, Daniyar Turmukhambetov, Gabriel J Brostow, and Michael Firman. Learning stereo from single images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 722–740. Springer, 2020. 3
- [110] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. In *ICCV*, 2023. 2
- [111] Hongyu Wen, Erich Liang, and Jia Deng. Layeredflow: A real-world benchmark for non-lambertian multi-layer optical flow. *arXiv preprint arXiv:2409.05688*, 2024. Accepted to ECCV 2024. 2, 6, 7
- [112] 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. 2, 4, 7, 8
- [113] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 7, 8
- [114] Peng Xu, Zhiyu Xiang, Chenyu Qiao, Jingyun Fu, and Xijun Zhao. Adaptive multi-modal cross-entropy loss for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [115] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019. 2
- [116] 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
- [117] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 1, 2, 3, 5, 6, 8, 18, 19, 31
- [118] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. 3
- [119] 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 *ICCV*, 2023. 3
- [120] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 3
- [121] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6044–6053, 2019. 2

- [122] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III* 14, pages 298–313. Springer, 2019. 3
- [123] Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Open challenges in deep stereo: the booster dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022. CVPR. 1, 2, 6
- [124] Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Luigi Di Stefano, Jean-Baptiste Weibel, Dominik Bauer, Doris Antensteiner, Markus Vincze, Jiaqi Li, Yachuan Huang, Junrui Zhang, Yiran Wang, Jinghong Zheng, Liao Shen, Zhiguo Cao, Ziyang Song, Zerong Wang, Ruijie Zhu, Hao Zhang, Rui Li, Jiang Wu, Xian Li, Yu Zhu, Jinqiu Sun, Yanning Zhang, Pihai Sun, Yuanqi Yao, Wenbo Zhao, Kui Jiang, Junjun Jiang, Mykola Lavreniuk, and Jui-Lin Wang. Tricky 2024 challenge on monocular depth from images of specular and transparent surfaces. In *European Conference on Computer Vision Workshops (ECCVW)*, 2024. 2
- [125] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, et al. NTIRE 2024 challenge on HR depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2
- [126] Jure Žbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1592–1599, 2015. 2
- [127] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(65):1–32, 2016. 2
- [128] Jiaxi Zeng, Chengtang Yao, Lidong Yu, Yuwei Wu, and Yunde Jia. Parameterized cost volume for stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18347–18357, 2023. 2, 7, 8
- [129] Jiaxi Zeng, Chengtang Yao, Yuwei Wu, and Yunde Jia. Temporally consistent stereo matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [130] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 185–194, 2019. 2
- [131] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *Europe Conference on Computer Vision (ECCV)*, 2020. 3
- [132] Youmin Zhang, Matteo Poggi, and Stefano Mattoccia. Temporalstereo: Efficient spatial-temporal stereo matching network. In *IROS*, 2023. 2
- [133] Yongjian Zhang, Longguang Wang, Kunhong Li, Yun Wang, and Yulan Guo. Learning representations from foundation models for domain generalized stereo matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [134] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 international conference on 3D vision (3DV)*, pages 668–678. IEEE, 2022. 3
- [135] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1327–1336, 2023. 2, 6, 7, 8, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30
- [136] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 3
- [137] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 36–53, 2018. 3

# Stereo Anywhere: Robust Zero-Shot Deep Stereo Matching Even Where Either Stereo or Mono Fail

## Supplementary Material

This document reports additional material concerning “Stereo Anywhere: Robust Zero-Shot Deep Stereo Matching Even Where Either Stereo or Mono Fail”. Specifically:

- First, we present an extended description of our proposed architecture in Sec. 3, including detailed formulations of the monocular correlation volume (Sec. 7.1), differentiable monocular scaling (Sec. 7.2), cost volume augmentation (Sec. 7.3), volume truncation (Sec. 7.4), and training supervision (Sec. 7.5).
- We then report extensive ablation studies in Sec. 8 demonstrating how our stereo matching architecture effectively generalizes across different state-of-the-art monocular depth networks (Sec. 8.1), showing consistent improvements over baseline stereo methods regardless of the specific VFM employed. Then, we show qualitatively the impact of the truncated cost volume augmentation on disparity estimation on non-Lambertian surfaces (Sec. 8.2). Furthermore, we include an analysis of runtime performance and memory consumption (Sec. 8.3) across different input resolutions and VFMs.
- Finally, we present extensive qualitative results in Sec. 9 across multiple datasets, demonstrating the effectiveness of our method in dealing with challenging scenarios such as non-Lambertian surfaces, transparent objects and textureless regions.

## 7. Method Overview: Additional Details

In this section, we enrich the description of Stereo Anywhere architecture.

### 7.1. Monocular Correlation Volume

Given the monocular depth estimations  $\mathbf{M}_L \in \mathbb{R}^{1 \times H \times W}$  and  $\mathbf{M}_R \in \mathbb{R}^{1 \times H \times W}$ , we aim to estimate the normal maps  $\nabla_L \in \mathbb{R}^{3 \times \frac{H}{4} \times \frac{W}{4}}$  and  $\nabla_R \in \mathbb{R}^{3 \times \frac{H}{4} \times \frac{W}{4}}$  to construct the 3D correlation volume  $\mathbf{V}_M \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times \frac{W}{4}}$ . We decide to use  $\nabla_L$  and  $\nabla_R$  instead of extracting additional features from  $\mathbf{M}_L$  and  $\mathbf{M}_R$  because  $\mathbf{M}_L$  and  $\mathbf{M}_R$  already provide high-level information. Furthermore, normal maps can handle depth inconsistencies between  $\mathbf{M}_L$  and  $\mathbf{M}_R$  that can occur for example when a foreground object is visible only in a single view. We downsample  $\mathbf{M}_L$  and  $\mathbf{M}_R$  to 1/4 – bilinear interpolation, then we estimate  $\nabla_L$  and  $\nabla_R$  – spatial gradient:

$$\nabla = \frac{\nabla^*}{\|\nabla^*\|}, \quad \nabla^* = \begin{bmatrix} \frac{\partial(\lambda \mathbf{M}_{\frac{1}{4}})}{\partial x} & \frac{\partial(\lambda \mathbf{M}_{\frac{1}{4}})}{\partial y} & 1 \end{bmatrix}, \quad \lambda = \frac{1}{10} \cdot \frac{W}{4} \quad (8)$$

where  $\lambda$  is a gain factor that is proportional to  $W$ , which permits to achieve scale-invariant normal maps.

Given the absence of texture in normal maps,  $\mathbf{V}_M$  will be not ambiguous only in edges. To alleviate this problem, we segment  $\mathbf{V}_M$  – the relative depth priors from  $\mathbf{M}_L$  and  $\mathbf{M}_R$ : doing so we aim to reduce the ambiguity by forcing the matching only in similar depth regions (e.g., foreground objects cannot match with background object since the correlation score is masked to zero). Considering Eq. (3), we calculate masks  $\mathcal{M}_L^n$  and  $\mathcal{M}_R^n$  as follows:

$$(\mathcal{M}_L^n)_{ij} = \begin{cases} 1 & \text{if } \frac{n}{N} \leq (\mathbf{M}_L)_{ij} < \frac{n+1}{N} \\ 0 & \text{otherwise} \end{cases} \quad (\mathcal{M}_R^n)_{ik} = \begin{cases} 1 & \text{if } \frac{n}{N} \leq (\mathbf{M}_R)_{ik} < \frac{n+1}{N} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

To further deal with the ambiguity, we improve the 3D Convolutional Regularization model  $\phi_A$  – an adapted version of CoEx [4] correlation volume excitation that exploits both views  $\mathbf{M}_L$ ,  $\mathbf{M}_R$ :

$$(\mathbf{V}'^s M)_{fijk} = \sigma((\mathbf{f}_L^s)_{fij}) \odot \sigma((\mathbf{f}_R^s)_{fik}) \odot (\mathbf{V}^s M)_{fijk} \quad (10)$$

where  $\mathbf{V}'^s M$  is the excited volume,  $\sigma(\cdot)$  is the sigmoid function,  $\odot$  is the element-wise product,  $\mathbf{V}^s M \in \mathbb{R}^{F \times \frac{H}{s} \times \frac{W}{s} \times \frac{W}{s}}$  is an intermediate correlation feature volume at scale  $s$  with  $F$  features inside module  $\phi_A$ ,  $\mathbf{f}_L^s \in \mathbb{R}^{F \times \frac{H}{s} \times \frac{W}{s} \times 1}$  and  $\mathbf{f}_R^s \in \mathbb{R}^{F \times \frac{H}{s} \times 1 \times \frac{W}{s}}$  are shallow 2D conv-features extracted from  $\mathbf{M}_L$  and  $\mathbf{M}_R$  downsampled at proper scale.

### 7.2. Differentiable Monocular Scaling

As detailed in Sec. 3.2, volume  $\mathbf{V}_M^D$  is used also to estimate the coarse disparity maps  $\hat{\mathbf{D}}_L$ ,  $\hat{\mathbf{D}}_R$ , while volume  $\mathbf{V}_M^C$  is utilized to estimate confidence maps  $\hat{\mathbf{C}}_L$ ,  $\hat{\mathbf{C}}_R$ .  $\hat{\mathbf{D}}_L$ ,  $\hat{\mathbf{C}}_L$  and  $\hat{\mathbf{D}}_R$ ,  $\hat{\mathbf{C}}_R$  are used to scale respectively  $\mathbf{M}_L$  and  $\mathbf{M}_R$ . As described in Eq. (4), we can estimate left disparity from a correlation volume – a softargmax operation on the last  $W$  dimension of  $\mathbf{V}_M^D$

and – the relationship between left disparity and correlation. Here we report an extended version of Eq. (4) with the explicit formula for softargmax operator:

$$(\hat{\mathbf{D}}_L)_{ij} = j - \left( \text{softargmax}_L(\mathbf{V}_M^D) \right)_{ij} = j - \sum_d^{\frac{W}{4}} d \cdot \frac{e^{(\mathbf{V}_M^D)_{ijd}}}{\sum_f^{\frac{W}{4}} e^{(\mathbf{V}_M^D)_{iff}}} \quad (11)$$

At the same time, given the relationship between right disparity and correlation  $d_R = k_L - k_R$  we can estimate the right disparity performing a softargmax on the first  $W$  dimension of  $\mathbf{V}_M^D$ :

$$(\hat{\mathbf{D}}_R)_{ik} = \left( \text{softargmax}_R(\mathbf{V}_M^D) \right)_{ik} - k = \sum_d^{\frac{W}{4}} d \cdot \frac{e^{(\mathbf{V}_M^D)_{idk}}}{\sum_f^{\frac{W}{4}} e^{(\mathbf{V}_M^D)_{ifk}}} - k \quad (12)$$

Disparity maps  $\hat{\mathbf{D}}_L$   $\hat{\mathbf{D}}_R$  are used in combination with confidence maps  $\hat{\mathbf{C}}_L$   $\hat{\mathbf{C}}_R$  to obtain a robust scaling. We present an expanded version of the information entropy based confidence estimation (Eq. (5)), with the explicit formula for softmax operator:

$$(\hat{\mathbf{C}}_L)_{ij} = 1 + \frac{\sum_d^{\frac{W}{4}} (\text{softmax}_L(\mathbf{V}_M^C))_{ijd} \cdot \log_2 ((\text{softmax}_L(\mathbf{V}_M^C))_{ijd})}{\log_2(\frac{W}{4})} = 1 + \frac{\sum_d^{\frac{W}{4}} \frac{e^{(\mathbf{V}_M^C)_{ijd}}}{\sum_f^{\frac{W}{4}} e^{(\mathbf{V}_M^C)_{iff}}} \cdot \log_2 \left( \frac{e^{(\mathbf{V}_M^C)_{ijd}}}{\sum_f^{\frac{W}{4}} e^{(\mathbf{V}_M^C)_{iff}}} \right)}{\log_2(\frac{W}{4})} \quad (13)$$

In the same way, we estimate right confidence map  $\hat{\mathbf{C}}_R$  performing a softmax operation on the first  $W$  dimension of  $\mathbf{V}_M^C$ :

$$(\hat{\mathbf{C}}_R)_{ik} = 1 + \frac{\sum_d^{\frac{W}{4}} (\text{softmax}_R(\mathbf{V}_M^C))_{idk} \cdot \log_2 ((\text{softmax}_R(\mathbf{V}_M^C))_{idk})}{\log_2(\frac{W}{4})} = 1 + \frac{\sum_d^{\frac{W}{4}} \frac{e^{(\mathbf{V}_M^C)_{idk}}}{\sum_f^{\frac{W}{4}} e^{(\mathbf{V}_M^C)_{ifk}}} \cdot \log_2 \left( \frac{e^{(\mathbf{V}_M^C)_{idk}}}{\sum_f^{\frac{W}{4}} e^{(\mathbf{V}_M^C)_{ifk}}} \right)}{\log_2(\frac{W}{4})} \quad (14)$$

To improve the robustness of the scaling, we introduce a softLRC operator to classify occlusions as low-confidence pixels and consequentially mask out them from  $\hat{\mathbf{C}}_L$  and  $\hat{\mathbf{C}}_R$ . We define the softLRC operator as follows:

$$\text{softLRC}_L(\mathbf{D}_L, \mathbf{D}_R) = \frac{\log(1 + \exp(T_{\text{LRC}} - |\mathbf{D}_L - \mathcal{W}_L(\mathbf{D}_L, \mathbf{D}_R)|))}{\log(1 + \exp(T_{\text{LRC}}))} \quad (15)$$

where  $T_{\text{LRC}} = 1$  is the LRC threshold and  $\mathcal{W}_L(\mathbf{D}_L, \mathbf{D}_R)$  is the warping operator that uses the left disparity  $\mathbf{D}_L$  to warp the right disparity  $\mathbf{D}_R$  into the left view.

Finally, we can estimate the scale  $\hat{s}$  and shift  $\hat{t}$  – a differentiable weighted least-square approach. We report here the expanded form of Eq. (6):

$$\min_{\hat{s}, \hat{t}} \left\| \sqrt{\hat{\mathbf{C}}_L} \odot [(\hat{s}\mathbf{M}_L + \hat{t}) - \hat{\mathbf{D}}_L] \right\|_F + \left\| \sqrt{\hat{\mathbf{C}}_R} \odot [(\hat{s}\mathbf{M}_R + \hat{t}) - \hat{\mathbf{D}}_R] \right\|_F \quad (16)$$

where  $\|\cdot\|_F$  denotes the Frobenius norm.

### 7.3. Cost Volume Augmentations

Volume augmentations are necessary when the training set – *e.g.*, Sceneflow [58] – does not model particularly complex scenarios where a VFM could be useful, for example, when experiencing non-Lambertian surfaces. Without any augmentation of this kind, the stereo network would simply overlook the additional information from the monocular branch. As detailed in the main paper, we propose three volume augmentations and a monocular augmentation to overcome this issue. In this supplementary section, we explain the rationale behind the introduction of each augmentation:

- *Volume Rolling*: non-Lambertian surfaces such as mirrors and glasses violate the geometry constraints, leading to a high matching peak in a wrong disparity bin. This augmentation emulates this behavior by shifting some among the matching peaks to a random position: consequentially, Stereo Anywhere learns to retrieve the correct peak from the other branch.
- *Volume Noising* and *Volume Zeroing*: we introduce noise and false peaks into the correlation volume to simulate scenarios with texture-less regions, repeating patterns, and occlusions.

- *Perfect Monocular Estimation*: instead of acting inside the correlation volumes, we can substitute the prediction of the VFM with a perfect monocular map – the ground truth normalized between [0, 1]. This perfect prediction is noise-free and therefore the monocular branch of Stereo Anywhere will likely gain importance during the training process.

## 7.4. Volume Truncation

The proposed volume truncation strategy further helps Stereo Anywhere to handle mirror surfaces. Here we introduce additional details about fuzzy operators – useful to make a boolean expression differentiable – and the sigmoid curve used to truncate the volume  $\mathbf{V}_S$  – the truncate mask  $(\mathbf{T})_{ij} = \left[ ((\hat{\mathbf{M}}_L)_{ij} > (\hat{\mathbf{D}}_L)_{ij}) \wedge (\mathbf{C}_M)_{ij} \right] \vee \left[ (\mathbf{C}_M)_{ij} \wedge \neg(\hat{\mathbf{C}}_L)_{ij} \right]$ .

We can replace operators AND ( $\wedge$ ), OR ( $\vee$ ), NOT ( $\neg$ ) and GREATER ( $>$ ) inside  $\mathbf{T}$  with the fuzzy counterparts  $\text{AND}_F(A, B) = A \cdot B$ ,  $\text{OR}_F(A, B) = A + B - A \cdot B$ ,  $\text{NOT}_F(A) = 1 - A$  and  $\text{GREATER}_F(A, B) = \sigma(A - B)$ , obtaining the fuzzy truncate mask  $\mathbf{T}_F$ :

$$\begin{aligned} (\mathbf{T}_F)_{ij} &= (\mathbf{T}_F^A)_{ij} + (\mathbf{T}_F^B)_{ij} - (\mathbf{T}_F^A)_{ij} \cdot (\mathbf{T}_F^B)_{ij} \\ (\mathbf{T}_F^A)_{ij} &= (\mathbf{C}_M)_{ij} \cdot \sigma((\hat{\mathbf{M}}_L)_{ij} - (\hat{\mathbf{D}}_L)_{ij}) \\ (\mathbf{T}_F^B)_{ij} &= (\mathbf{C}_M)_{ij} \cdot (1 - (\hat{\mathbf{C}}_L)_{ij}) \end{aligned} \quad (17)$$

where  $\mathbf{T}_F^A$  and  $\mathbf{T}_F^B$  are respectively the left section and the right section of the  $\text{OR}_F$  of mask  $\mathbf{T}_F$ . Next, we can apply threshold  $T_m$  to achieve the final fuzzy mask  $\mathbf{T}'_F$  as follows:

$$(\mathbf{T}'_F)_{ij} = \sigma((\mathbf{T}_F)_{ij} - T_m) \quad (18)$$

Finally, we can use the fuzzy truncate mask  $\mathbf{T}'_F$  and the scaled monocular map  $\hat{\mathbf{M}}_L$  to generate the sigmoid-based truncation volume  $\mathbf{V}_T$ :

$$(\mathbf{V}_T)_{ijk} = (1 - (\mathbf{T}'_F)_{ij}) + (\mathbf{T}'_F)_{ij} \cdot \left[ \sigma(j - (\hat{\mathbf{M}}_L)_{ij} - k) \cdot (1 - G) + G \right] \quad (19)$$

where  $G = 0.9$  attenuates the impact of the truncation. The correlation volume  $\mathbf{V}_S$  is truncated through an element-wise product with  $\mathbf{V}_T$ .

## 7.5. Training Supervision

We supervise the iterative module – the L1 loss with exponentially increasing weights [52]:

$$\mathcal{L}_A = \sum_{l=1}^L \gamma^{L-l} \|\mathbf{D}^l - \mathbf{D}_{\text{Lgt}}\|_1 \quad (20)$$

where  $L$  is the total number of iterations made by the update operator and  $\mathbf{D}_{\text{Lgt}}$  is the ground truth of the left disparity map. Furthermore, we supervise the outputs  $\hat{\mathbf{D}}_L, \hat{\mathbf{D}}_R, \hat{\mathbf{M}}_L, \hat{\mathbf{M}}_R, \hat{\mathbf{C}}_L, \hat{\mathbf{C}}_R$  of the monocular branch – respectively L1 loss and normal loss for  $\hat{\mathbf{D}}_L, \hat{\mathbf{D}}_R$ , L1 loss for  $\hat{\mathbf{M}}_L, \hat{\mathbf{M}}_R$  and Binary Cross Entropy (BCE) loss for  $\hat{\mathbf{C}}_L, \hat{\mathbf{C}}_R$ :

$$\mathcal{L}_B = \|\hat{\mathbf{D}}_L - \mathbf{D}_{\text{Lgt}}\|_1 + \psi \left\| \mathbf{1} - \nabla_L \cdot \hat{\nabla}_L \right\|_1 \quad \left( \nabla_L \cdot \hat{\nabla}_L \right)_{ij} = \sum_h (\nabla_L)_{hij} \cdot (\hat{\nabla}_L)_{hij} \quad (21)$$

$$\mathcal{L}_C = \|\hat{\mathbf{D}}_R - \mathbf{D}_{\text{Rgt}}\|_1 + \psi \left\| \mathbf{1} - \nabla_R \cdot \hat{\nabla}_R \right\|_1 \quad \left( \nabla_R \cdot \hat{\nabla}_R \right)_{ik} = \sum_h (\nabla_R)_{hik} \cdot (\hat{\nabla}_R)_{hik} \quad (22)$$

$$\mathcal{L}_D = \|\hat{\mathbf{M}}_L - \mathbf{D}_{\text{Lgt}}\|_1 \quad \mathcal{L}_E = \|\hat{\mathbf{M}}_R - \mathbf{D}_{\text{Rgt}}\|_1 \quad (23)$$

$$\mathcal{L}_F = \text{BCE}(\hat{\mathbf{C}}_L, \mathbf{C}_{\text{Lgt}}), \quad (\mathbf{C}_{\text{Lgt}})_{ij} = \frac{\log \left( 1 + \exp \left( T_{\text{LRC}} - |(\hat{\mathbf{D}}_L)_{ij} - (\mathbf{D}_{\text{Lgt}})_{ij}| \right) \right)}{\log(1 + \exp(T_{\text{LRC}}))} \quad (24)$$

$$\mathcal{L}_G = \text{BCE}(\hat{\mathbf{C}}_R, \mathbf{C}_{\text{Rgt}}), \quad (\mathbf{C}_{\text{Rgt}})_{ik} = \frac{\log \left( 1 + \exp \left( T_{\text{LRC}} - |(\hat{\mathbf{D}}_R)_{ik} - (\mathbf{D}_{\text{Rgt}})_{ik}| \right) \right)}{\log(1 + \exp(T_{\text{LRC}}))} \quad (25)$$

where  $\psi = 10$  is the normal loss weight,  $\mathbf{D}_{\text{Rgt}}$  is the ground truth of the right disparity map,  $\hat{\nabla}_L, \hat{\nabla}_R$  are the normal maps estimated respectively from  $\hat{\mathbf{D}}_L, \hat{\mathbf{D}}_R$ ,  $\nabla_L \cdot \hat{\nabla}_L$  and  $\nabla_R \cdot \hat{\nabla}_R$  are the dot product between normal maps, and  $\mathbf{C}_{\text{Lgt}}, \mathbf{C}_{\text{Rgt}}$  are the confidence ground truth. The final supervision loss  $\mathcal{L}$  is the sum of all previous partial losses:

$$\mathcal{L} = \mathcal{L}_A + \mathcal{L}_B + \mathcal{L}_C + \mathcal{L}_D + \mathcal{L}_E + \mathcal{L}_F + \mathcal{L}_G \quad (26)$$

## 8. Additional Ablation Studies

In this section, we report additional studies concerning the performance of Stereo Anywhere.

### 8.1. Generalization to Different VFM

In the main paper, we assumed Depth Anything v2 [117] as the VFM fueling Stereo Anywhere, since it is the latest state-of-the-art model being published at the time of this submission. However, any VFM for monocular depth estimation would be suitable for this purpose, either current or future ones. To confirm this argument, we conducted some experiments by replacing Depth Anything v2 with other VFM that appeared on arXiv in the last months, yet that are not been officially published. Among them, we select DepthPro [6], MoGe [106] and Lotus [32].

Table 5 shows the results achieved by Stereo Anywhere variants – different VFM on Booster and LayeredFlow. We can appreciate how the different flavors of Stereo Anywhere always outperform the baseline stereo model [52]. In general, Depth Anything v2 remains the best choice to deal with non-Lambertian surfaces, with Moge allowing for small improvements on some metrics over the Booster dataset.

Model	Booster (Q)					LayeredFlow (E)				
	Error Rate (%)			Avg. (px)	> 1	Error Rate (%)			Avg. (px)	
	> 2	> 4	> 6			> 8	> 3	> 5		
Baseline [52]	17.84	13.06	10.76	9.24	3.59	89.21	79.02	71.61	19.27	
Stereo Anywhere – DAv2 [117]	9.96	5.81	4.48	3.79	1.36	80.83	58.21	46.48	12.14	
Stereo Anywhere – DepthPro [6]	10.53	7.02	5.79	5.13	2.40	78.76	61.11	51.04	14.43	
Stereo Anywhere – MoGe [106]	9.47	5.77	4.49	3.84	1.44	84.27	68.67	58.89	16.22	
Stereo Anywhere – Lotus [32]	12.44	8.71	7.58	6.98	3.21	86.04	62.75	49.47	13.98	

Table 5. Non-Lambertian Generalization of Stereo Anywhere w.r.t VFM. We measure the impact of different monocular depth estimation networks. Networks trained on SceneFlow [58].

Table 6 shows the results achieved by the different VFM on the zero-shot generalization benchmark. Also in this case, we can appreciate how any Stereo Anywhere variant yields comparable accuracy, with some VFM like Moge yielding some improvements over Depth Anything v2 on ETH3D, KITTI 2012 and 2015 at the expense of lowering the accuracy on Middlebury 2014 and 2021. Interestingly, we can observe an important drop in accuracy by using DepthPro on Middlebury 2021, due to several failures by the model itself on the scenes of this dataset.

Model	Middlebury 2014 (H)			Middlebury 2021			ETH3D			KITTI 2012			KITTI 2015			
	bad > 2		Avg. (px)	bad > 2		Avg. (px)	bad > 1		Avg. (px)	bad > 3		Avg. (px)	bad > 3		Avg. (px)	
	All	Noc	Occ	All	Noc	Occ	All	Noc	Occ	All	Noc	Occ	All	Noc	Occ	
Baseline [52]	11.15	8.06	29.06	1.55	12.05	9.38	37.89	1.81	2.59	2.24	8.78	0.25	4.80	3.70	28.54	0.89
Stereo Anywhere – DAv2 [117]	7.07	4.76	20.77	0.97	8.38	5.86	32.87	1.10	2.39	2.16	5.82	0.28	3.94	3.03	21.02	0.85
Stereo Anywhere – DepthPro [6]	6.58	4.32	20.05	0.99	15.13	12.52	41.16	8.97	2.74	2.54	6.09	0.44	3.13	2.25	18.25	0.75
Stereo Anywhere – MoGe [106]	7.79	5.23	22.86	1.21	9.86	7.30	33.48	1.28	1.28	1.09	3.78	0.21	2.85	2.00	17.40	0.73
Stereo Anywhere – Lotus [32]	7.35	4.96	21.71	1.07	9.62	7.01	34.92	1.29	2.68	2.44	6.04	0.31	4.54	3.58	22.71	0.92

Table 6. Generalization of Stereo Anywhere w.r.t VFM. We measure the impact of different monocular depth estimation networks. Networks trained on SceneFlow [58].

Finally, Figure 7 shows qualitative results obtained by the different variants of Stereo Anywhere, highlighting only minor differences among the different predictions.

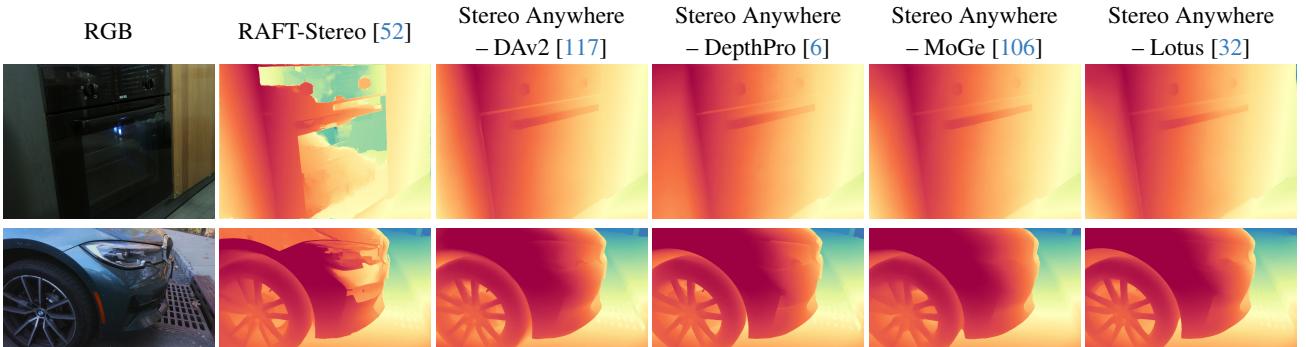


Figure 7. Qualitative Results – Booster and LayeredFlow. Predictions by RAFT-Stereo and Stereo Anywhere – different VFM.

## 8.2. Impact of Cost Volume Truncation

Cost volume truncation is a specific augmentation we apply to improve the results in the presence of mirrors. Figure 8 shows a qualitative example of predictions by Stereo Anywhere (using Depth Anything v2) obtained by either not applying or by applying such augmentation. While Stereo Anywhere alone cannot entirely restore the surface of the mirror starting from the priors provided by the VFM, applying cost volume truncation allows for predicting a much smoother and consistent surface.

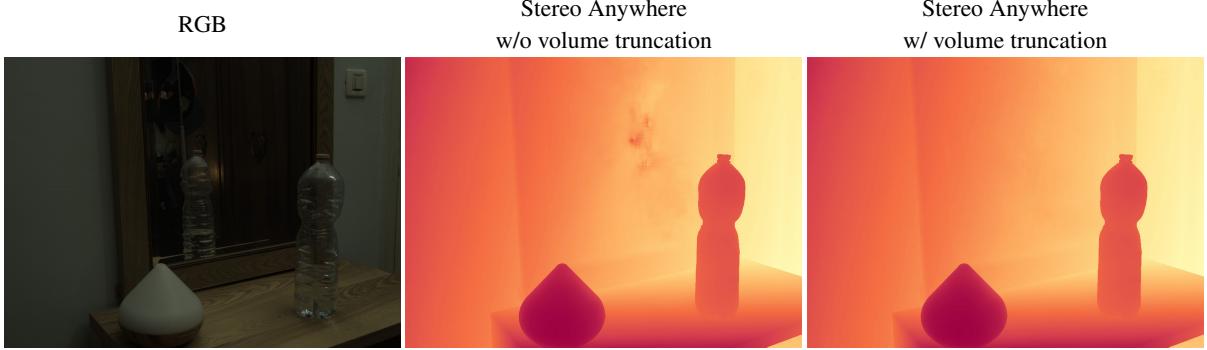


Figure 8. **Qualitative Results – Volume Truncation.** Predictions by Stereo Anywhere.

## 8.3. Runtime & Memory Consumption Analysis

Table 7 reports the processing time (in seconds) and memory consumption (in GB) required by Stereo Anywhere during inference, comparing it with the baseline stereo backbone, RAFT-Stereo. We measure the runtime on a single A100 GPU, repeating the experiment with three different input resolutions, specifically  $256 \times 256$ ,  $512 \times 512$ , and  $1024 \times 1024$ , as well as by deploying the different VFMs studied before to fuel Stereo Anywhere – specifically, for each variant we report standalone runtime and memory usage by the VFM and the stereo backbone separately, as well as their sum.

Concerning runtime, Depth Anything v2 is the fastest among the VFMs, taking about 30ms to process a single image at any resolution, with Moge requiring more than  $10\times$  the time for a single inference when processing 1Mpx images. The stereo backbone requires about 50% additional time compared to the baseline, RAFT-Stereo [52], because of the additional branch deployed to process the depth maps by the VFM.

For what concerns memory consumption, once again Depth Anything v2 is the most efficient among the VFMs, requiring as few as 2GB, with Moge sharing similar requirements. Our stereo backbone introduces additional memory consumption because of the second branch processing monocular cues: this overhead is negligible with 256 images, raising to about  $2\times$  the memory required by RAFT-Stereo alone when dealing with 1Mpx images.

Image Size ( $H \times W$ )	Stereo Model Name	VFM Name	Processing Time (s)			Memory Consumption (GB)		
			VFM	Stereo	Total	VFM	Stereo	Total
$256 \times 256$	<b>Stereo Anywhere (ours)</b>	DAv2 [117]	0.03	0.15	0.18	0.57	0.18	0.76
		DepthPro [6]	0.21	0.15	0.36	1.92	0.18	2.09
		MoGe [106]	0.38	0.15	0.52	0.38	0.19	0.57
		Lotus [32]	0.13	0.15	0.29	0.22	0.18	0.41
$256 \times 256$	RAFT-Stereo [52]	-	-	0.10	0.10	-	0.17	0.17
$512 \times 512$	<b>Stereo Anywhere (ours)</b>	DAv2 [117]	0.03	0.21	0.24	0.57	0.77	1.34
		DepthPro [6]	0.20	0.21	0.41	1.84	0.77	2.60
		MoGe [106]	0.38	0.21	0.59	0.38	0.78	1.17
		Lotus [32]	0.16	0.22	0.38	0.85	0.77	1.62
$512 \times 512$	RAFT-Stereo [52]	-	-	0.14	0.14	-	0.66	0.66
$1024 \times 1024$	<b>Stereo Anywhere (ours)</b>	DAv2 [117]	0.03	0.61	0.63	0.58	5.73	6.31
		DepthPro [6]	0.21	0.61	0.82	1.85	5.73	7.59
		MoGe [106]	0.38	0.60	0.98	0.42	5.77	6.19
		Lotus [32]	0.49	0.61	1.10	3.40	5.73	9.13
$1024 \times 1024$	RAFT-Stereo [52]	-	-	0.36	0.36	-	2.63	2.63

Table 7. **Runtime & Memory Consumption Analysis.**

## 9. Qualitative Results

We conclude with additional qualitative results by Stereo Anywhere on the different datasets involved in our experiments.

Figure 9 shows two examples from the KITTI 2012 dataset (respectively, stereo pairs 000040 and 000068). We can notice how any existing stereo model is unable to properly perceive the presence of transparent surfaces, as in correspondence of the windows on buildings and cars. On the contrary Stereo Anywhere, driven by the priors injected through the VFM, properly predicts the disparity corresponding to the transparent surfaces.

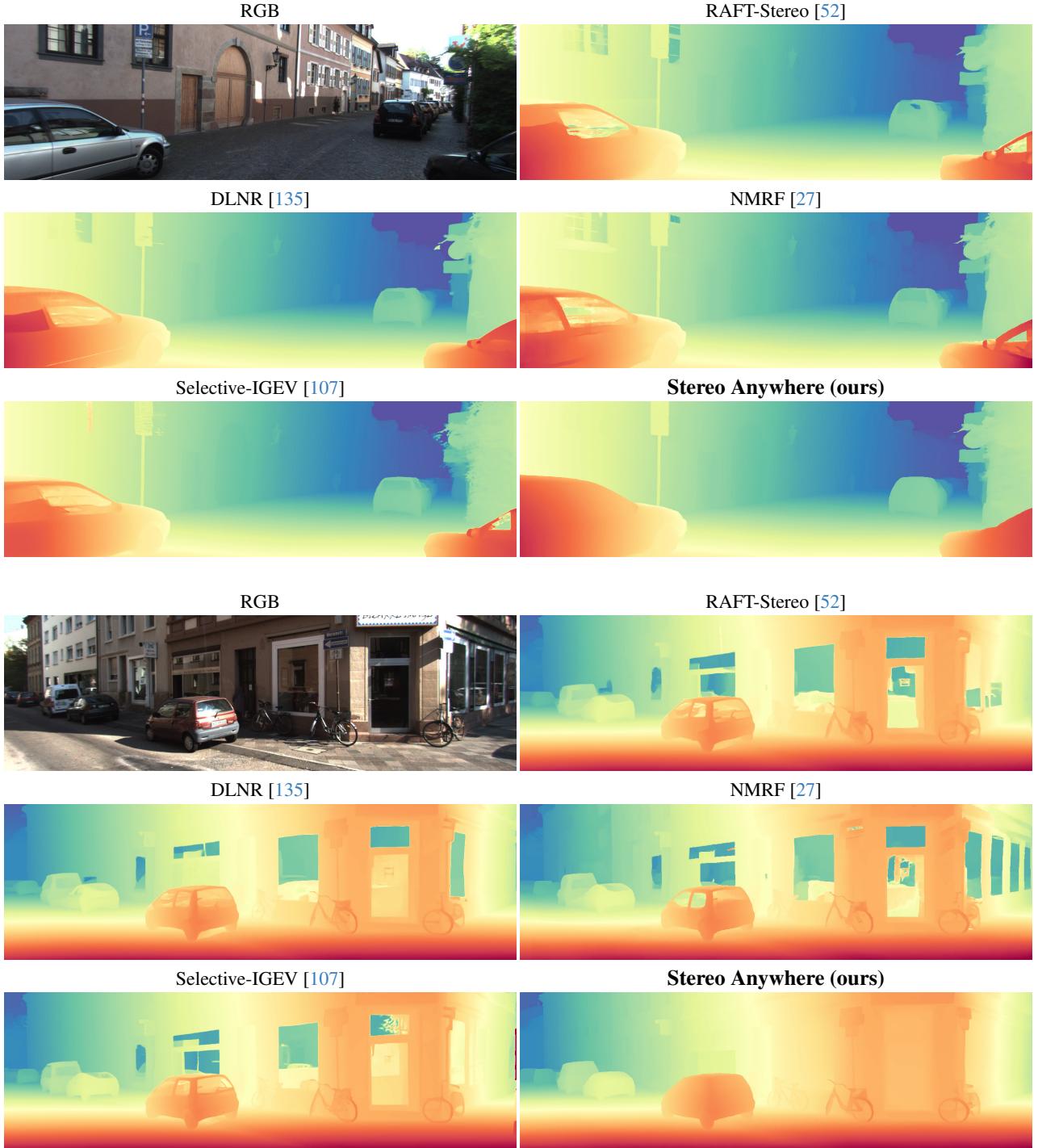


Figure 9. **Qualitative Results – KITTI 2012 (part 1).** Predictions by state-of-the-art models and Stereo Anywhere.

Figure 10 shows two further examples from KITTI 2012 (respectively, stereo pairs 000073 and 000127). In this case, we can appreciate the much higher level of detail in the disparity maps predicted by Stereo Anywhere, with extremely thin structures in fences and gates being preserved.

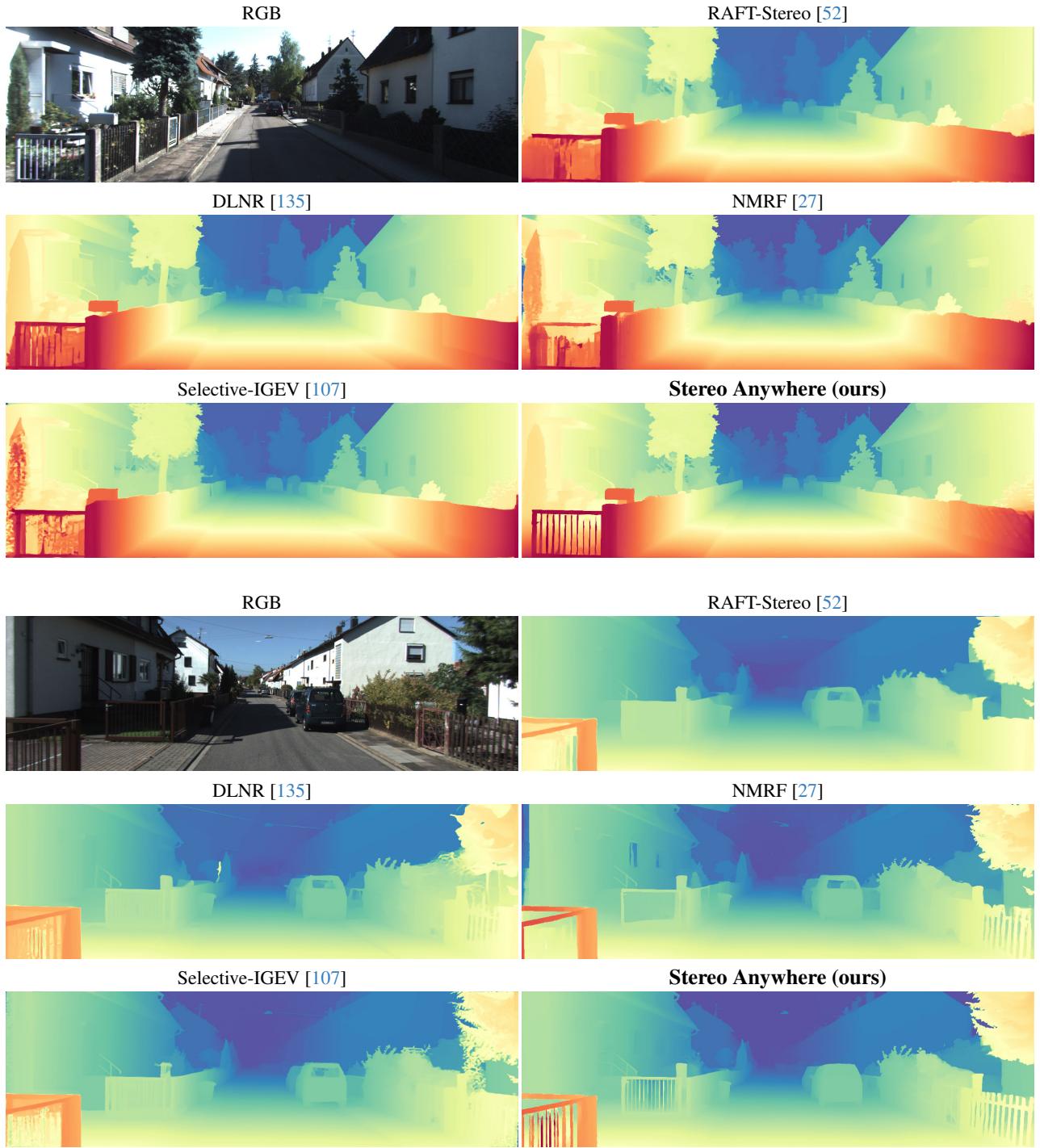


Figure 10. **Qualitative Results – KITTI 2012 (part 2).** Predictions by state-of-the-art models and Stereo Anywhere.

Figure 11 reports two stereo pairs from KITTI 2015 (respectively, 000024 and 000049). These examples confirm the ability to recover both thin structures and transparent surfaces already appreciated in KITTI 2012.

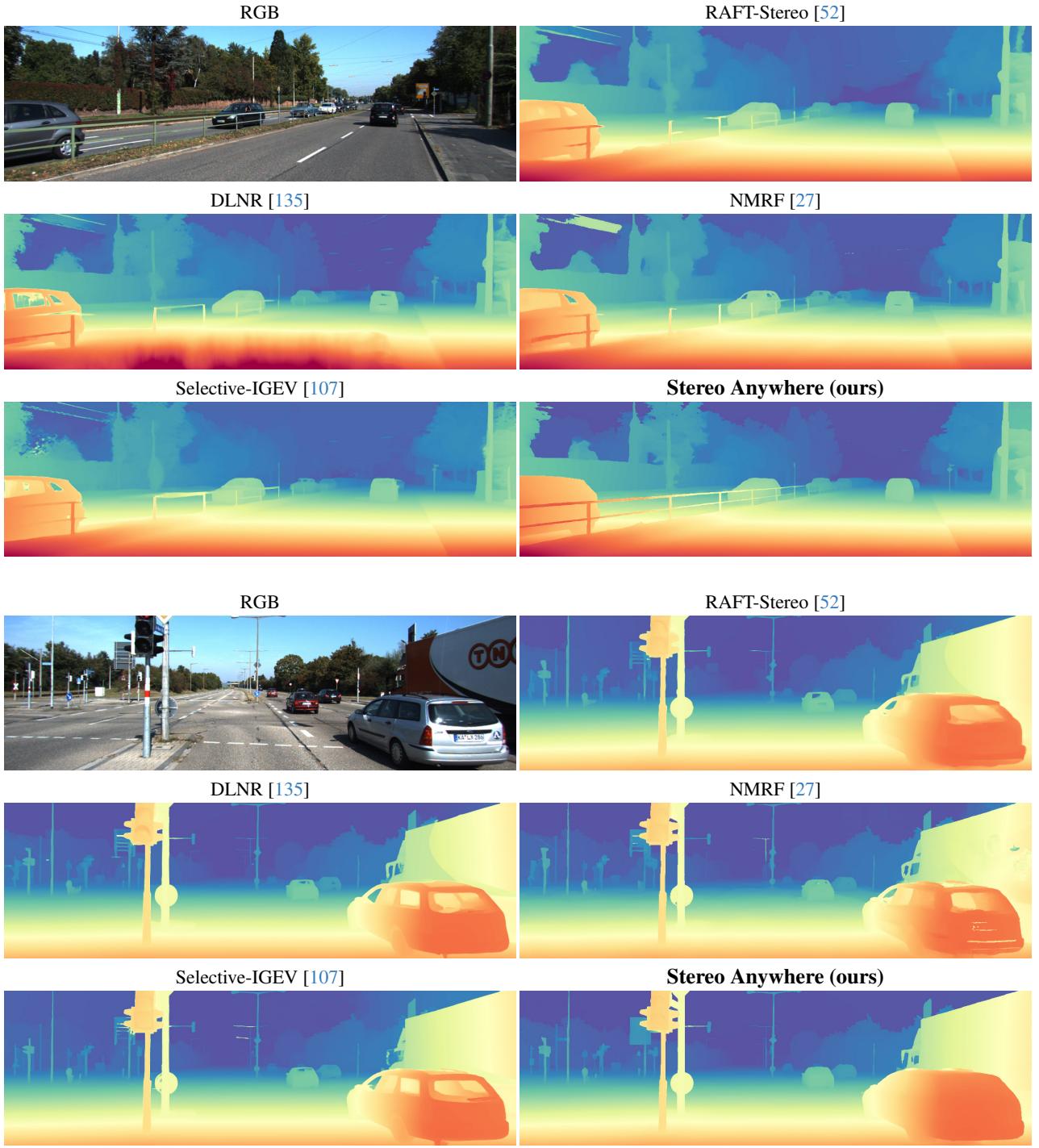


Figure 11. **Qualitative Results – KITTI 2015 (part 1).** Predictions by state-of-the-art models and Stereo Anywhere.

Figure 12 reports two additional samples from KITTI 2015 (respectively, 000093 and 000144). These latter present both underexposed and transparent regions, respectively on the billboard and the tram in the two images. While existing stereo networks struggle at dealing with both, Stereo Anywhere exposes unprecedented robustness.

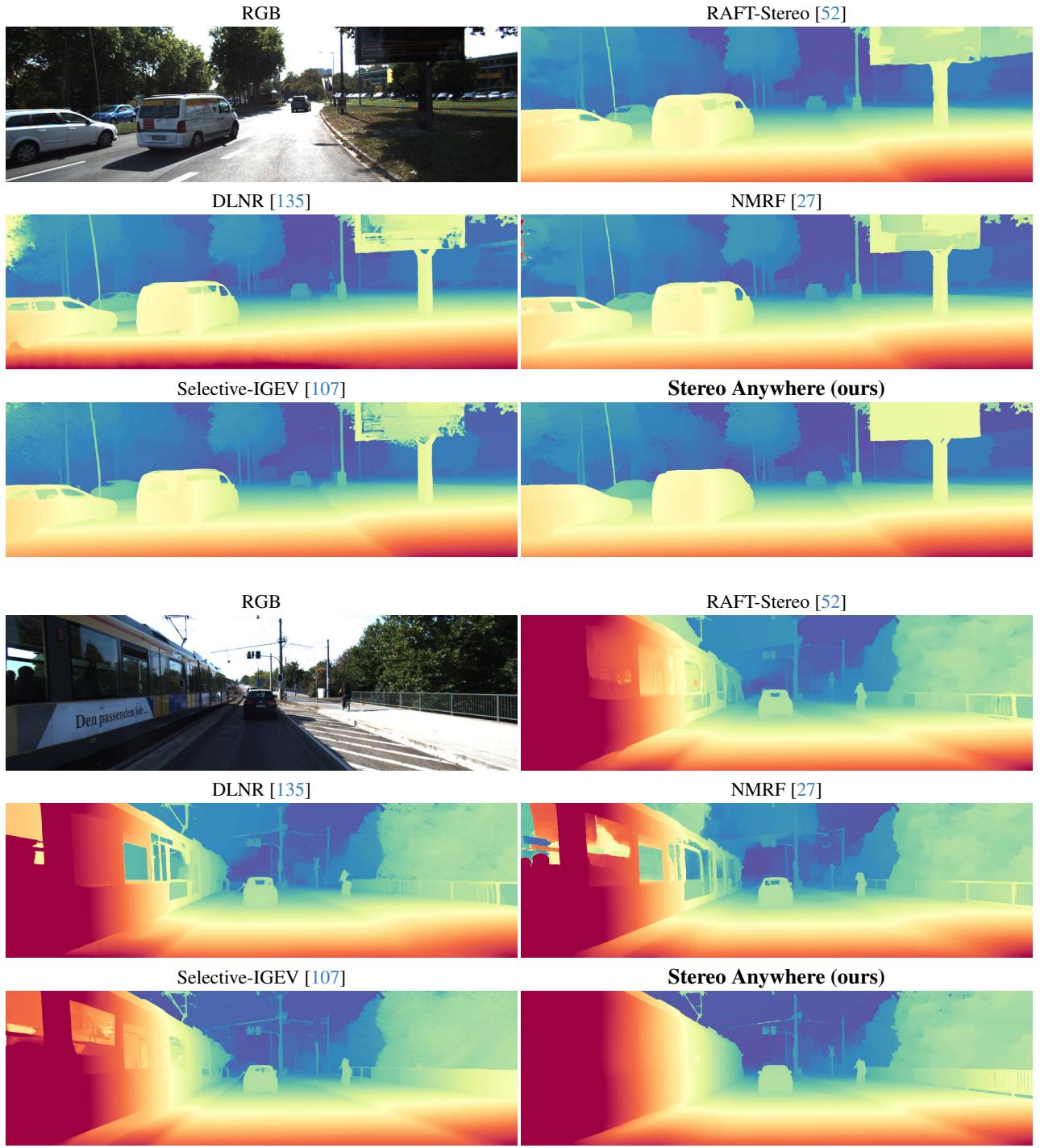


Figure 12. **Qualitative Results – KITTI 2015 (part 2).** Predictions by state-of-the-art models and Stereo Anywhere.

Figure 13 reports two image pairs from Middlebury 2014 (respectively, *Adirondack* and *Vintage*). On the former, Stereo Anywhere preserves the very thin holes on the back of the chair, while on the latter it can properly estimate the disparity for the displays, where existing methods are fooled and predict holes.

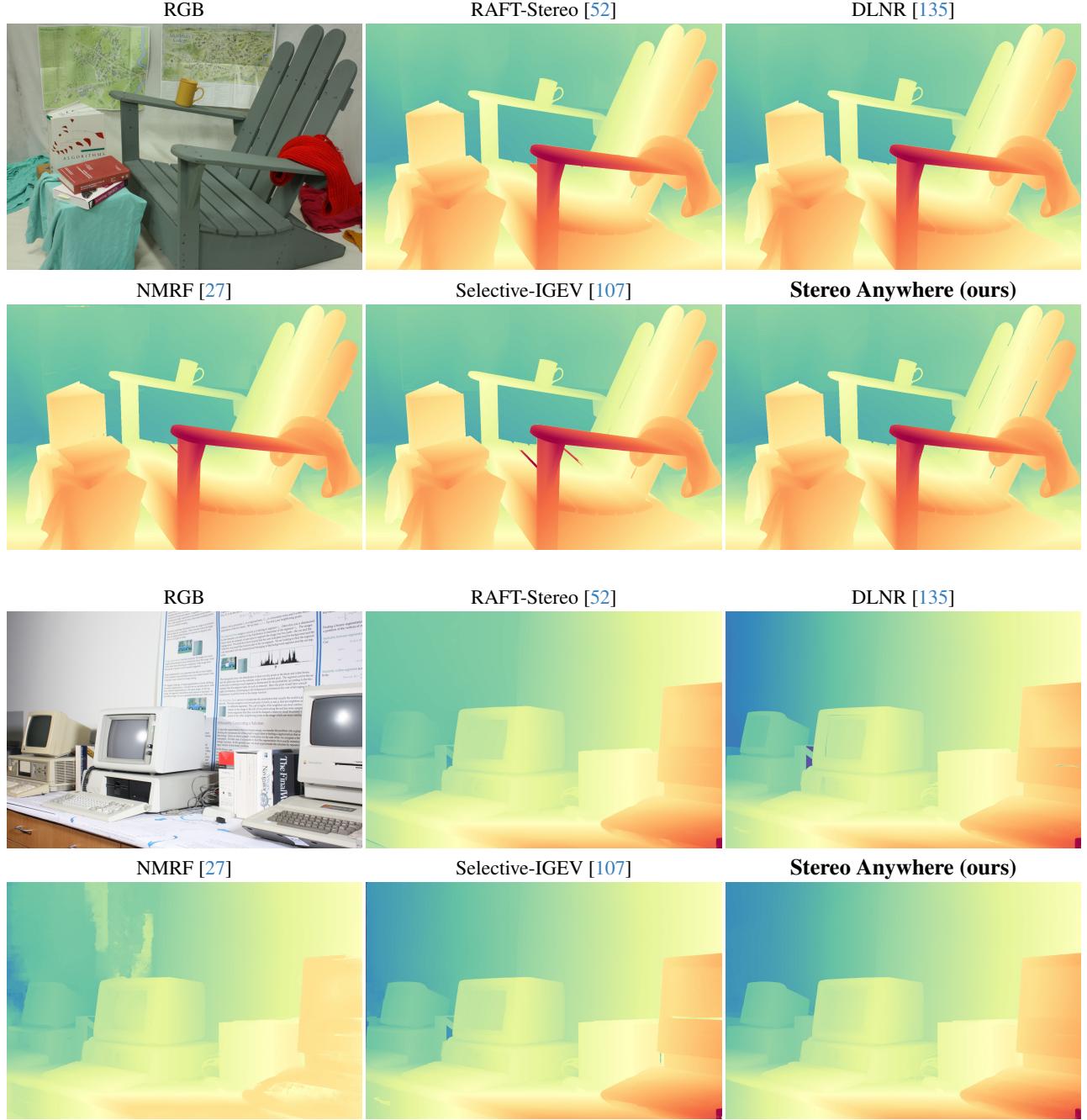


Figure 13. **Qualitative Results – Middlebury 2014.** Predictions by state-of-the-art models and Stereo Anywhere.

Figure 14 and 15 shows the results on two samples from Middlebury 2021, peculiar for their aspect ratio (respectively, *ladder1* and *ladder2*). Although existing models perform quite well on both, they fail to preserve the skittles on the top of the scene, whereas Stereo Anywhere properly predicts their structure.

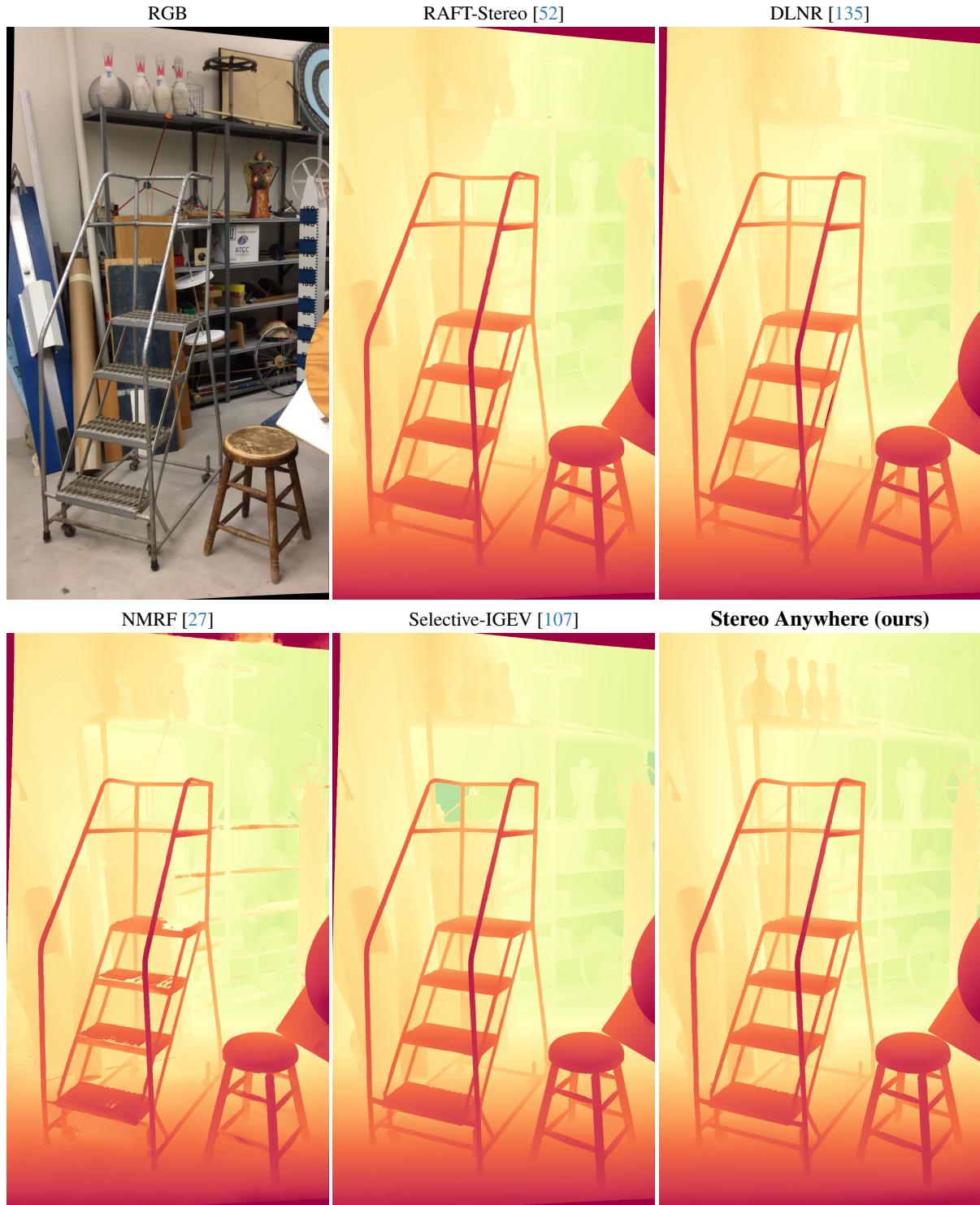


Figure 14. Qualitative Results – Middlebury 2021 (part 1). Predictions by state-of-the-art models and Stereo Anywhere.



Figure 15. Qualitative Results – Middlebury 2021 (part 2). Predictions by state-of-the-art models and Stereo Anywhere.

Figure 16 collects three outdoor images from ETH3D (respectively, *Playground1*, *Playground2* and *Playground3*). Once again, Stereo Anywhere proves its supremacy at predicting fine details such as branches and poles, while resulting more robust to challenging illumination conditions such as the sun flare in *Playground2*.

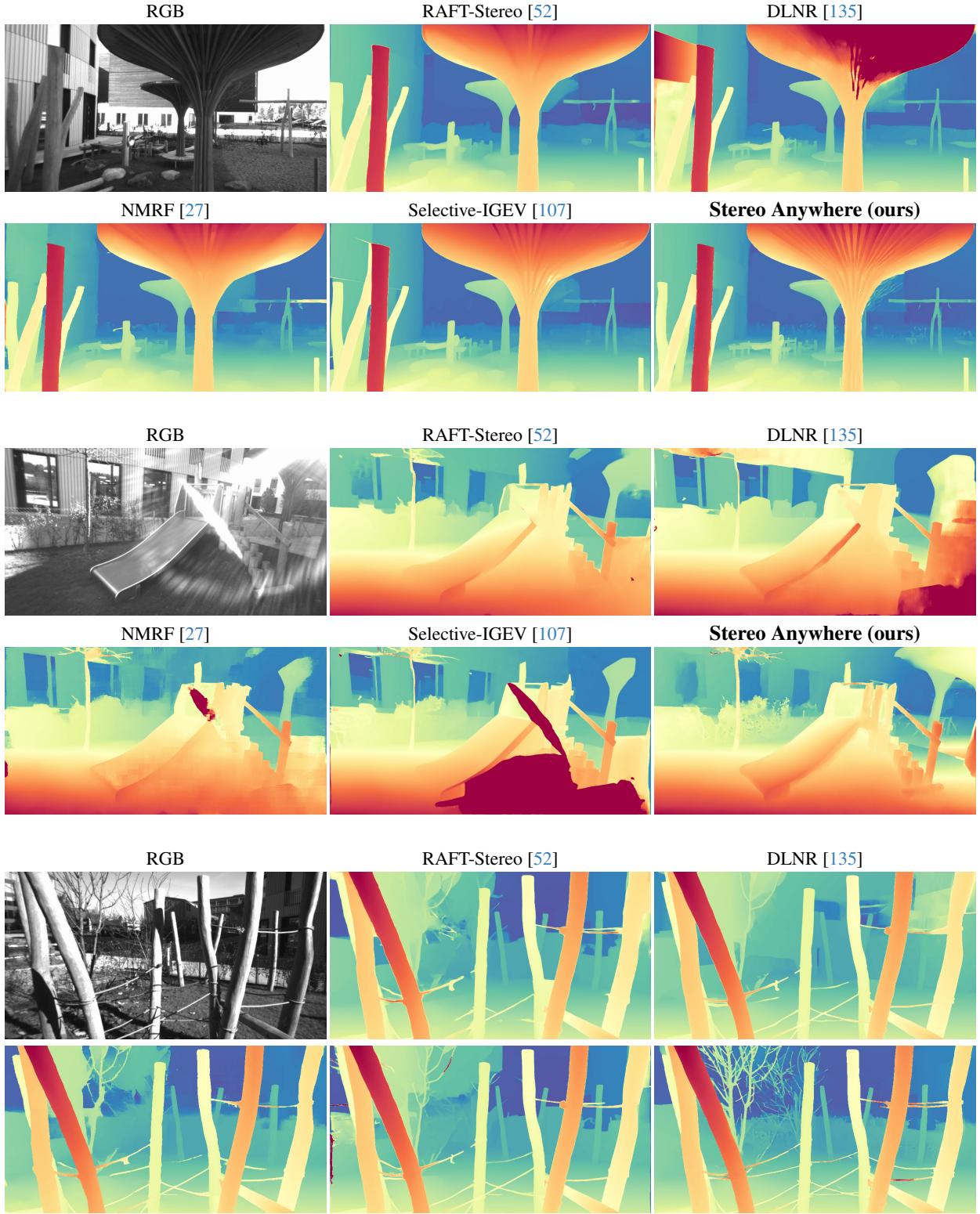


Figure 16. **Qualitative Results – ETH3D.** Predictions by state-of-the-art models and Stereo Anywhere.

Figure 17 and 18 report four examples from the Booster dataset, confirming how Stereo Anywhere can exploit the strong priors provided by the VFM to properly perceive the glass surface on the window in the former image, as well as challenging, untextured black surfaces of the computer, the TV and the displays appearing in the remaining samples.

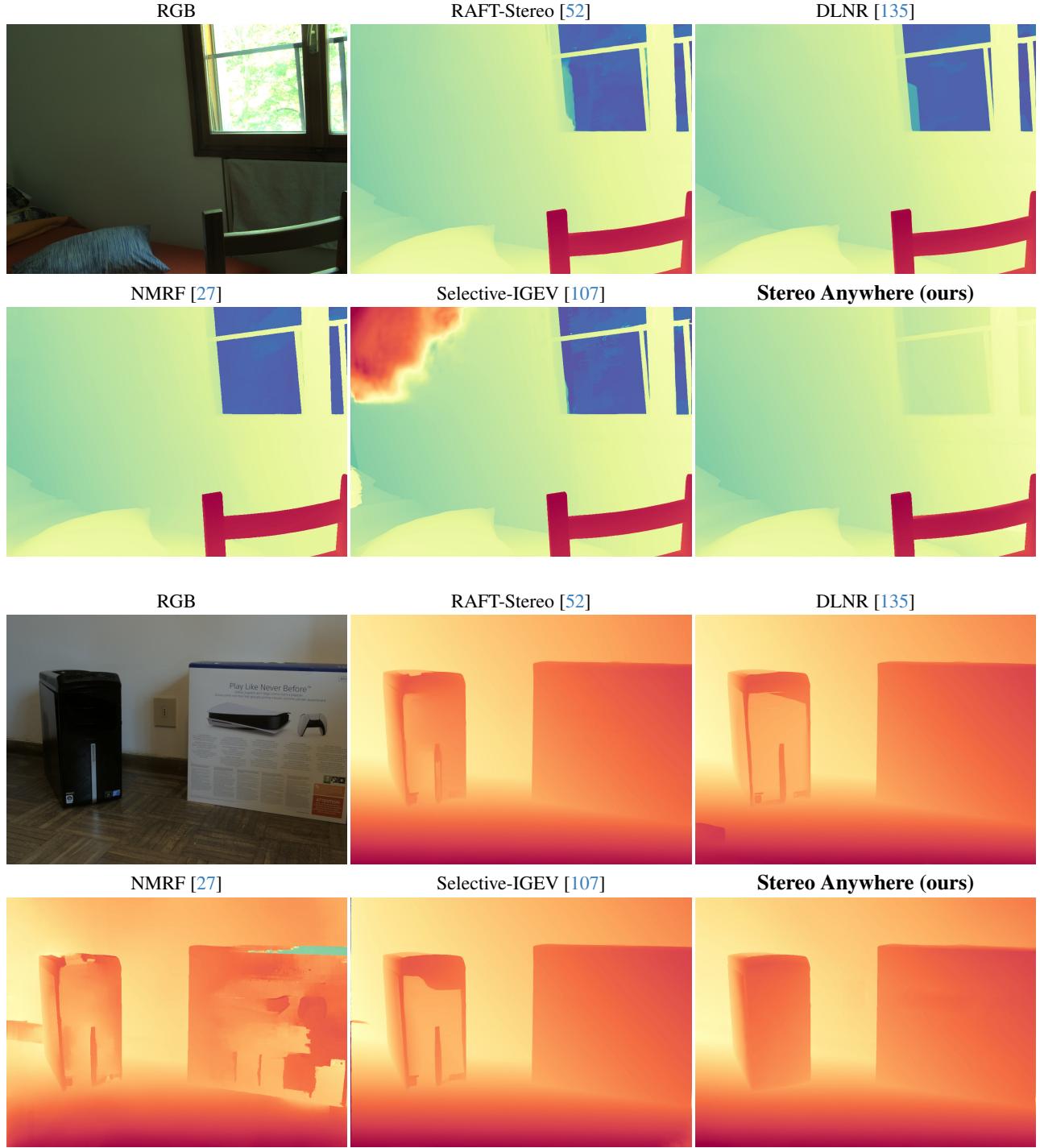


Figure 17. **Qualitative Results – Booster (part 1).** Predictions by state-of-the-art models and Stereo Anywhere.

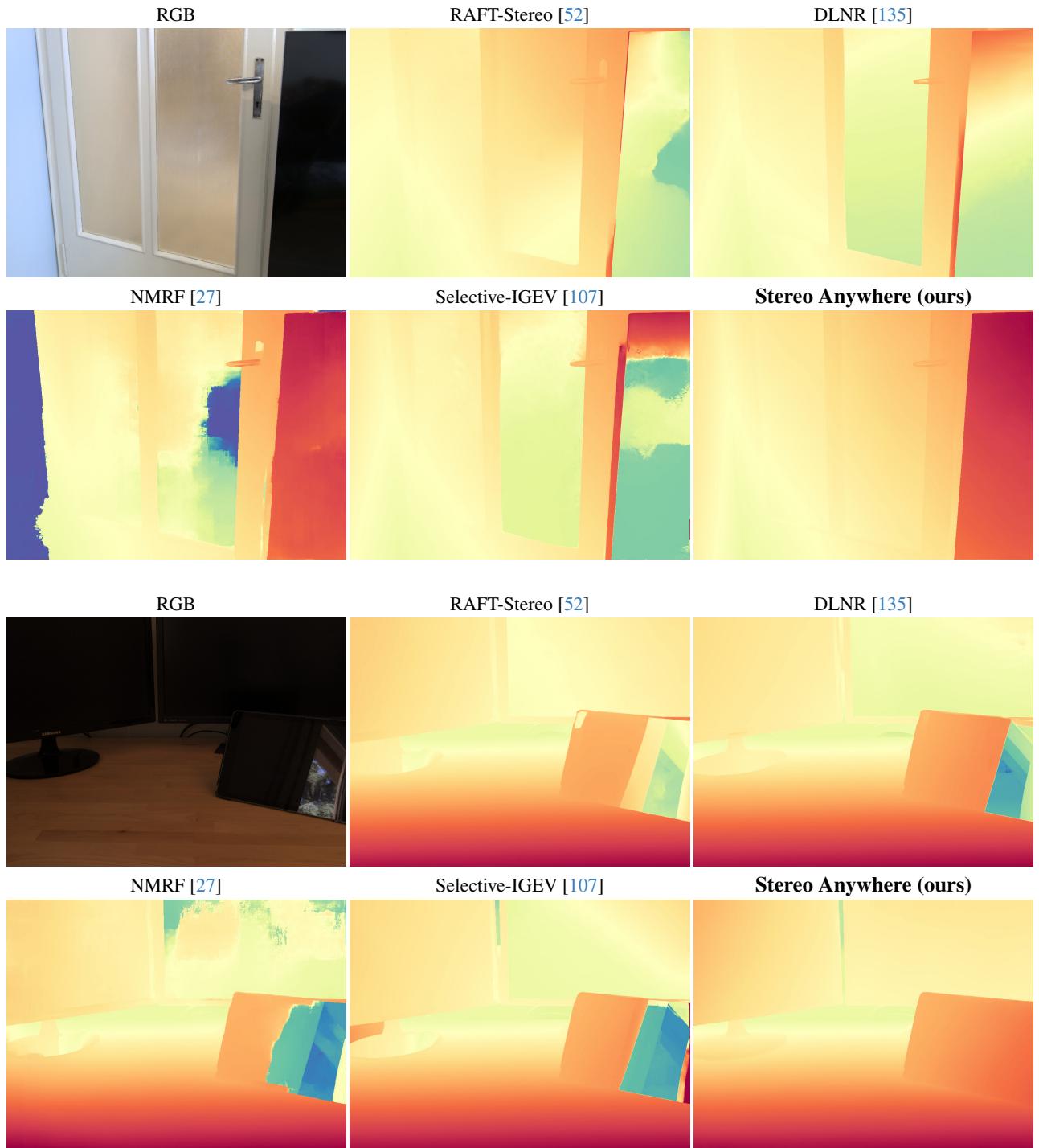


Figure 18. **Qualitative Results – Booster (part 2).** Predictions by state-of-the-art models and Stereo Anywhere.

Figure 19 showcases three images from the LayeredFlow dataset, highlighting once again the inability of the state-of-the-art networks to model even small, transparent surfaces as those in the doors from the first and second samples, conversely to Stereo Anywhere which can properly identify their presence. Finally, the third sample further highlights the high level of detail in Stereo Anywhere predictions once again.

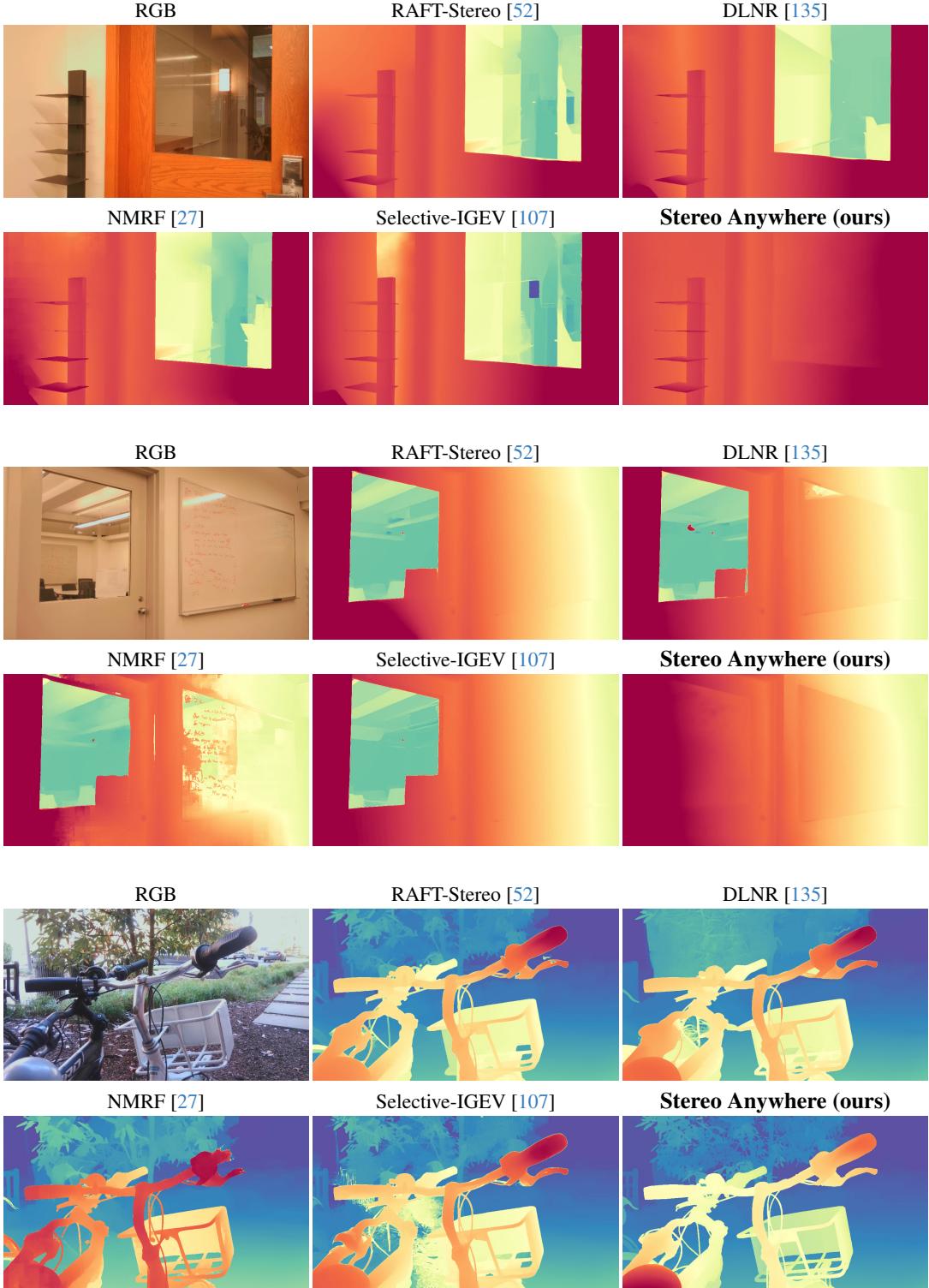


Figure 19. **Qualitative Results – LayeredFlow.** Predictions by state-of-the-art models and Stereo Anywhere.

To conclude, Figure 20 collects three scenes from our novel MonoTrap dataset. In this case, we report predictions by both state-of-the-art monocular and stereo models, as well as by Stereo Anywhere. The perspective illusions fooling monocular methods, unsurprisingly, do not affect stereo networks, which however are inaccurate near the left border of the image (first sample) or in the absence of texture (second sample). Stereo Anywhere effectively combines the strength of both worlds, while being not affected by any of their weaknesses.

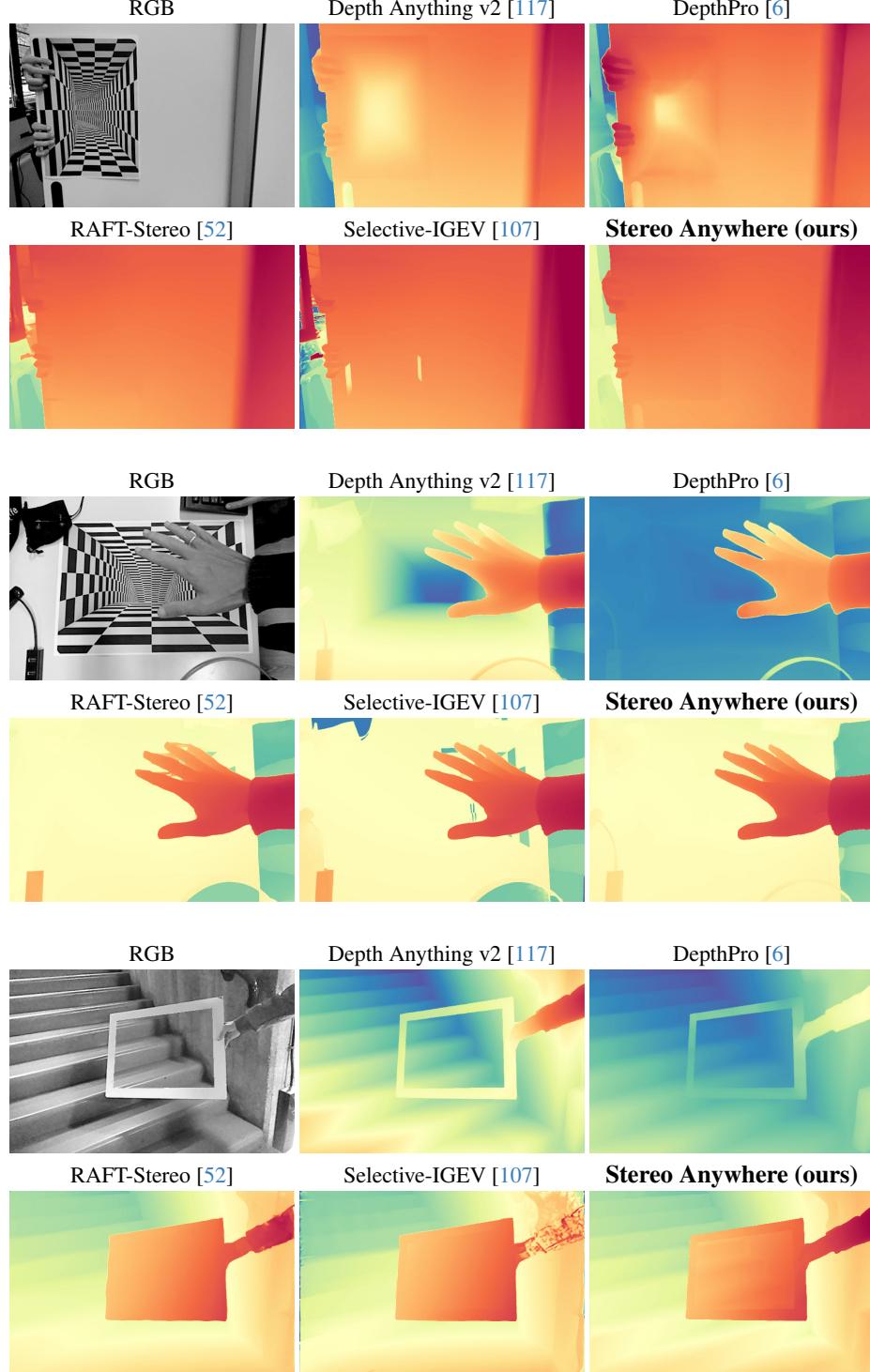


Figure 20. **Qualitative Results – MonoTrap.** Predictions by state-of-the-art models and Stereo Anywhere.