FoundationStereo: Zero-Shot Stereo Matching

Bowen Wen Matthew Trepte Joseph Aribido Jan Kautz
Orazio Gallo Stan Birchfield

NVIDIA

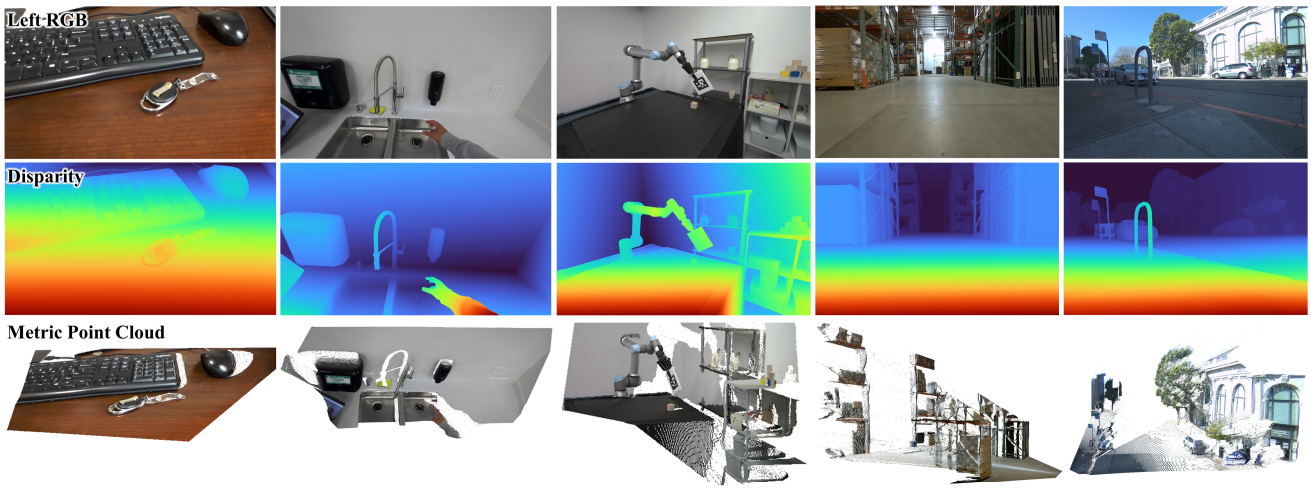

Figure 1. Zero-shot prediction on in-the-wild images. Our method generalizes to diverse scenarios (indoor / outdoor), objects of challenging properties (textureless / reflective / translucent / thin-structured), complex illuminations (shadow / strong exposure), various viewing perspectives and sensing ranges.

Abstract

Tremendous progress has been made in deep stereo matching to excel on benchmark datasets through per-domain fine-tuning. However, achieving strong zero-shot generalization — a hallmark of foundation models in other computer vision tasks — remains challenging for stereo matching. We introduce FoundationStereo, a foundation model for stereo depth estimation designed to achieve strong zero-shot generalization. To this end, we first construct a large-scale (1M stereo pairs) synthetic training dataset featuring large diversity and high photorealism, followed by an automatic self-curation pipeline to remove ambiguous samples. We then design a number of network architecture components to enhance scalability, including a side-tuning feature backbone that adapts rich monocular priors from vision foundation models to mitigate the sim-to-real gap, and long-range context reasoning for effective cost volume filtering. Together, these components lead to strong robustness and accuracy across domains, establishing a new standard in zero-shot stereo depth estimation. Project page: https://nvlabs.github.io/FoundationStereo/

1. Introduction

Since the advent of the first stereo matching algorithm nearly half a century ago [39], we have come a long way. Recent stereo algorithms can achieve amazing results, almost saturating the most challenging benchmarks—thanks to the proliferation of training datasets and advances in deep neural network architectures. Yet, fine-tuning on the dataset of the target domain is *still* the method of choice to get competitive results. Given the zero-shot generalization ability shown on other problems within computer vision via the scaling law [29, 42, 73, 74], what prevents stereo matching algorithms from achieving a similar level of generalization?

Leading stereo networks [9, 38, 49, 50, 68, 75] construct cost volumes from the unary features and leverage 3D CNNs for cost filtering. Refinement-based methods [12, 19, 24, 31, 33, 55, 62, 81] iteratively refine the disparity map based on recurrent modules such as Gated Recurrent Units (GRU). Despite their success on public benchmarks under per-domain fine-tuning setup, however, they struggle to gather non-local information to effectively scale to larger datasets. Other methods [32, 63] explore transformer architectures for unary feature extraction, while lacking the specialized structure afforded by cost volumes and iterative refinement to achieve high accuracy.

Such limitations have, to date, hindered the development of a stereo network that generalizes well to other domains. While it is true that cross-domain generalization has been explored by some prior works [8, 15, 34, 45, 77, 79], such

approaches have not achieved results that are competitive with those obtained by fine-tuning on the target domain, either due to insufficient structure in the network architecture, impoverished training data, or both. These networks are generally experimented on Scene Flow [40], a rather small dataset with only 40K annotated training image pairs. As a result, none of these methods can be used as an off-the-shelf solution, as opposed to the strong generalizability of vision foundation models that have emerged in other tasks.

To address these limitations, we propose FoundationStereo, a large foundation model for stereo depth estimation that achieves strong zero-shot generalization without per-domain fine-tuning. We train the network on a large-scale (1M image pairs) high-fidelity synthetic training dataset with high diversity and photorealism. An automatic self-curation pipeline is developed to eliminate the ambiguous samples that are inevitably introduced during the domain randomized data generation process, improving both the dataset quality and model robustness over iterate updates. To mitigate the sim-to-real gap, we propose a side-tuning feature backbone that adapts internet-scale rich priors from DepthAnythingV2 [74] that is trained on real monocular images to the stereo setup. To effectively leverage these rich monocular priors embedded into the 4D cost volume, we then propose an Attentive Hybrid Cost Volume (AHCF) module, consisting of 3D Axial-Planar Convolution (APC) filtering that decouples standard 3D convolution into two separate spatial- and disparity-oriented 3D convolutions, enhancing the receptive fields for volume feature aggregation; and a Disparity Transformer (DT) that performs self-attention over the entire disparity space within the cost volume, providing long range context for global reasoning. Together, these innovations significantly enhance the representation, leading to better disparity initialization, as well as more powerful features for the subsequent iterative refinement process.

Our contributions can be summarized as follows:

- We present FoundationStereo, a zero-shot generalizable stereo matching model that achieves comparable or even more favorable results to prior works fine-tuned on a target domain; it also significantly outperforms existing methods when applied to in-the-wild data.
- We create a large-scale (1M) high-fidelity synthetic dataset for stereo learning with high diversity and photorealism; and a self-curation pipeline to ensure that bad samples are pruned.
- To harness internet-scale knowledge containing rich semantic and geometric priors, we propose a Side-Tuning Adapter (STA) that adapts the ViT-based monocular depth estimation model [74] to the stereo setup.
- We develop Attentive Hybrid Cost Filtering (AHCF), which includes an hourglass module with 3D Axial-Planar Convolution (APC), and a Disparity Transformer (DT) module that performs full self-attention over the disparity dimension.

Code, model and dataset will be released.

2. Related Work

Deep Stereo Matching. Recent advances in stereo matching have been driven by deep learning, significantly enhancing accuracy and generalization. Cost volume aggregation methods construct cost volumes from unary features and perform 3D CNN for volume filtering [9, 38, 49, 50, 68, 75], though the high memory consumption prevents direct application to high resolution images. Iterative refinement methods, inspired by RAFT [53], bypasses the costly 4D volume construction and filtering by recurrently refining the disparity [12, 19, 24, 31, 33, 55, 62, 81]. While they generalize well to various disparity range, the recurrent updates are often time-consuming, and lack long-range context reasoning. Recent works [66, 67] thus combine the strengths of cost filtering and iterative refinement. With the tremendous progress made by vision transformers, another line of research [21, 32, 63] introduces transformer architecture to stereo matching, particularly in the unary feature extraction stage. Despite their success on per-domain fine-tuning setup, zero-shot generalization still remains challenging. To tackle this problem, [8, 15, 34, 45, 77, 79] explore learning domain-invariant features for cross-domain generalization, with a focus on training on Scene Flow [40] dataset. However, the strong generalizability of vision foundation models emerged in other tasks that is supported by scaling law has yet to be fully realized in stereo matching for practical applications.

Stereo Matching Training Data. Training data is essential for deep learning models. KITTI 12 [18] and KITTI 15 [41] provide hundreds of training pairs on driving scenarios. DrivingStereo [71] further scales up to 180K stereo pairs. Nevertheless, the sparse ground-truth disparity obtained by LiDAR sensors hinders learning accurate and dense stereo matching. Middlebury [47] and ETH3D [48] develop a low number of training data covering both indoor and outdoor scenarios beyond driving. Booster [44] presents a real-world dataset focusing on transparent objects. InStereo2K [1] presents a larger training dataset consisting of 2K stereo pairs with denser ground-truth disparity obtained with structured light system. However, challenges of scarce data size, imperfect ground-truth disparity and lack of collection scalability in real-world have driven the widespread adoption of synthetic data for training. This includes Scene Flow [40], Sintel [4], CREStereo [31], IRS [59], TartanAir [61], FallingThings [56], Virtual KITTI 2 [5], CARLA HR-VS [70], Dynamic Replica [25]. In Tab. 1, we compare our proposed FoundationStereo dataset (FSD) with commonly used synthetic training datasets for stereo matching. Our dataset encompasses a wide range of

		Sintel [4]	Scene Flow [40]	CREStereo [31]	IRS [59]	TartanAir [61]	FallingThings [56]	FSD (Ours)
Scenarios	Flying Objects	✗	✓	✓	✗	✗	✗	✓
	Indoor	✗	✗	✗	✓	✓	✓	✓
	Outdoor	✗	✓	✗	✗	✓	✓	✓
	Driving	✗	✓	✗	✗	✗	✗	✓
	Movie	✓	✓	✗	✗	✗	✗	✗
Simulator		Blender	Blender	Blender	Unreal Engine	Unreal Engine	Unreal Engine	NVIDIA Omniverse
Rendering Realism		High	Low	High	High	High	High	High
Scenes		10	9	0	4	18	3	12
Layout Realism		Medium	Low	Low	High	High	Medium	High
Stereo Pairs		1K†	40K†	200K	103K†	306K†	62K	1000K
Resolution		1024×436	960×540	1920×1080	960×540	640×480	960×540	1280×720
Reflections		✗	✗	✓	✓	✓	✓	✓
Camera Params		Constant	Constant	Constant	Constant	Constant	Constant	Varying‡

Table 1. Synthetic datasets for training stereo algorithms (excluding test images with inaccessible ground truth). †Indicates reduced diversity, caused by including many similar frames from video sequences. ‡Our dataset includes varying intrinsics and baseline.

scenarios, features the largest data volume to date, includes diverse 3D assets, captures stereo images under diversely randomized camera parameters, and achieves high fidelity in both rendering and spatial layouts.

Vision Foundation Models. Vision foundation models have significantly advanced across various vision tasks in 2D, 3D and multi-modal alignment. CLIP [43] leverages large-scale image-text pair training to align visual and textual modalities, enabling zero-shot classification and facilitating cross-modal applications. DINO series [6, 35, 42] employ self-supervised learning for dense representation learning, effectively capturing detailed features critical for segmentation and recognition tasks. SAM series [29, 46, 72] demonstrate high versatility in segmentation driven by various prompts such as points, bounding boxes, language. Similar advancements also appear in 3D vision tasks. DUSt3R [60] and MASt3R [30] present generalizable frameworks for dense 3D reconstruction from uncalibrated and unposed cameras. FoundationPose [64] develops a unified framework of 6D object pose estimation and tracking for novel objects. More closely related to this work, a number of efforts [2, 26, 73, 74] demonstrated strong generalization in monocular depth estimation task. Together, these approaches exemplify under the scaling law, how foundation models in vision are evolving to support robust applications across diverse scenarios without tedious per-domain fine-tuning.

3. Approach

The overall network architecture is shown in Fig. 2. The rest of this section describes the various components.

3.1. Monocular Foundation Model Adaptation

To mitigate the sim-to-real gap when the stereo network is primarily trained on synthetic dataset, we leverage the recent advancements on monocular depth estimation trained on internet-scale real data [3, 74]. We use a CNN network to adapt the ViT-based monocular depth estimation network to the stereo setup, thus synergizing the strengths of both CNN and ViT architectures.

We explored multiple design choices for combining CNN and ViT approaches, as outlined in Fig. 3 (left). In particular, (a) directly uses the feature pyramids from the DPT head in a frozen DepthAnythingV2 [74] without using CNN features. (b) resembles ViT-Adapter [10] by exchanging features between CNN and ViT. (c) applies a 4×4 convolution with stride 4 to downscale the feature before the DepthAnythingV2 final output head. The feature is then concatenated with the same level CNN feature to obtain a hybrid feature at 1/4 scale. The side CNN network is thus learned to adapt the ViT features [78] to stereo matching task. Surprisingly, while being simple, we found (c) significantly surpasses the alternative choices on the stereo matching task, as shown in the experiments (Sec. 4.5). As a result, we adopt (c) as the main design of STA module.

Formally, given a pair of left and right images $I_l, I_r \in \mathbb{R}^{H \times W \times 3}$, we employ EdgeNeXt-S [37] as the CNN module within STA to extract multi-level pyramid features, where the 1/4 level feature is equipped with DepthAnythingV2 feature: $f_l^{(i)}, f_r^{(i)} \in \mathbb{R}^{C_i \times \frac{H}{i} \times \frac{W}{i}}$, $i \in \{4, 8, 16, 32\}$. EdgeNeXt-S [37] is chosen for its memory efficiency and because larger CNN backbones did not yield additional benefits in our investigation. When forwarding to DepthAnythingV2, we first resize the image to be divisible by 14, to be consistent with its pretrained patch size. The STA weights are shared when applied to I_l, I_r.

Similarly, we employ STA to extract context feature, with the difference that the CNN module is designed with a sequence of residual blocks [23] and down-sampling layers. It generates context features of multiple scales: $f_c^{(i)} \in \mathbb{R}^{C_i \times \frac{H}{i} \times \frac{W}{i}}$, $i \in \{4, 8, 16\}$, as in [33]. f_c participates in initializing the hidden state of the ConvGRU block and inputting to the ConvGRU block at each iteration, effectively guiding the iterative process with progressively refined contextual information.

Fig. 3 visualizes the power of rich monocular prior that helps to reliably predict on ambiguous regions which is

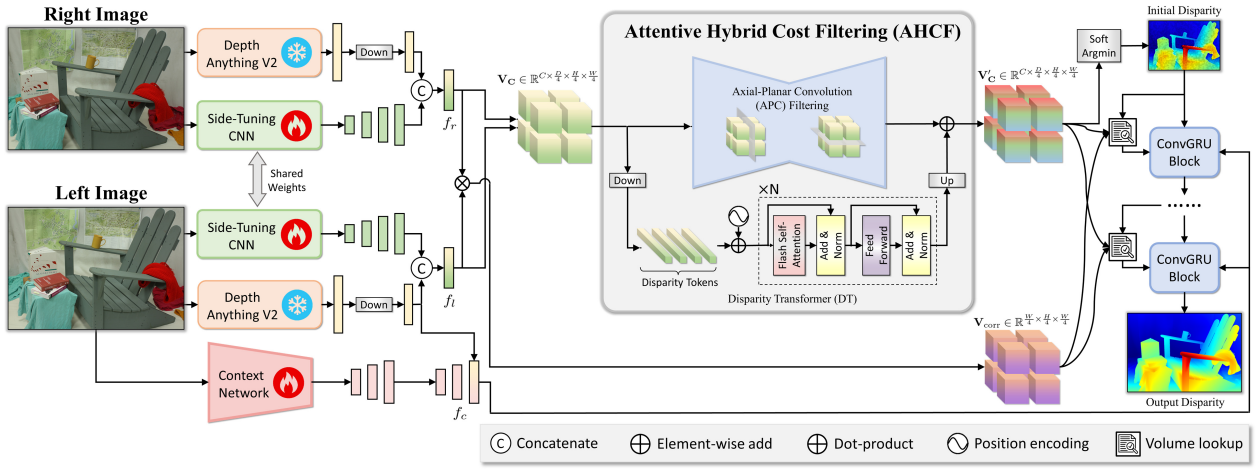

Figure 2. Overview of our proposed FoundationStereo. The Side-Tuning Adapter (STA) adapts the rich monocular priors from a frozen DepthAnythingV2 [74], while combined with fine-grained high-frequency features from multi-level CNN for unary feature extraction. Attentive Hybrid Cost Filtering (AHCF) combines the strengths of the Axial-Planar Convolution (APC) filtering and a Disparity Transformer (DT) module to effectively aggregate the features along spatial and disparity dimensions over the 4D hybrid cost volume. An initial disparity is then predicted from the filtered cost volume, and subsequently refined through GRU blocks. At each refinement step, the latest disparity is used to look up features from both filtered hybrid cost volume and correlation volume to guide the next refinement. The iteratively refined disparity becomes the final output.

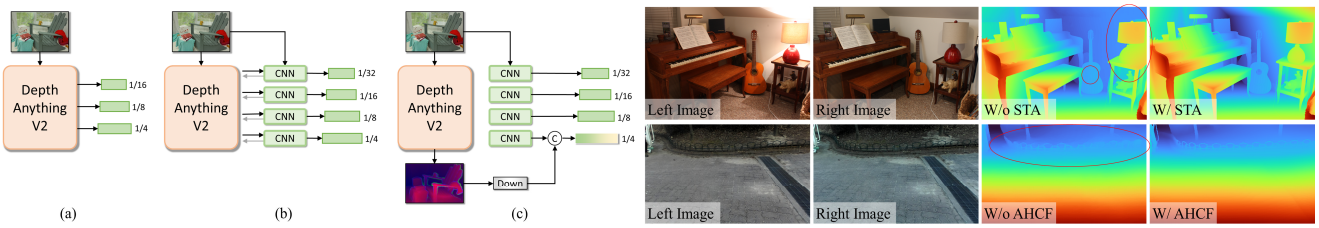

Figure 3. **Left:** Design choices for STA module. **Right:** Effects of the proposed STA and AHCF modules. "W/o STA" only uses CNN to extract features. "W/o AHCF" uses conventional 3D CNN-based hourglass network for cost volume filtering. Results are obtained via zero-shot inference without fine-tuning on target dataset. STA leverages rich monocular prior to reliably predict the lamp region with inconsistent lighting and dark guitar sound hole. AHCF effectively aggregates the spatial and long-range disparity context to accurately predict over thin repetitive structures.

challenging to deal with by naive correspondence search along the epipolar line.

3.2. Attentive Hybrid Cost Filtering

Hybrid Cost Volume Construction. Given unary features at 1/4 scale f_l^4, f_r^4 extracted from previous step, we construct the cost volume $\mathbf{V_C} \in \mathbb{R}^{C\times\frac{D}{4}\times\frac{H}{4}\times\frac{W}{4}}$ with a combination of group-wise correlation and concatenation [22]:

$$\mathbf{V}_{\text{gwc}}(g,d,h,w) = \left\langle \widehat{f}_{l,g}^{(4)}(h,w), \widehat{f}_{r,g}^{(4)}(h,w-d) \right\rangle,$$
$$\mathbf{V}_{\text{cat}}(d,h,w) = \left[\text{Conv}(f_l^{(4)})(h,w), \text{Conv}(f_r^{(4)})(h,w-d)\right],$$
$$\mathbf{V_C}(d,h,w) = [\mathbf{V}_{\text{gwc}}(d,h,w), \mathbf{V}_{\text{cat}}(d,h,w)] \quad (1)$$

where $\widehat{f}$ denotes L_2 normalized feature for better training stability; $\langle\cdot,\cdot\rangle$ represents dot product; $g \in \{1,2,...,G\}$ is the group index among the total $G = 8$ feature groups that we evenly divide the total features into; $d \in \{1,2,...,\frac{D}{4}\}$ is the disparity index. $[\cdot,\cdot]$ denotes concatenation along channel dimension. The group-wise correlation $\mathbf{V}_{\text{gwc}}$ harnesses the strengths of conventional correlation-based matching costs, offering a diverse set of similarity measurement features from each group. $\mathbf{V}_{\text{cat}}$ preserves unary features including the rich monocular priors by concatenating left and right features at shifted disparity. To reduce memory consumption, we linearly downsize the unary feature dimension to 14 using a convolution of kernel size 1 (weights are shared between f_l^4 and f_r^4) before concatenation. Next, we describe two sub-modules for effective cost volume filtering.

Axial-Planar Convolution (APC) Filtering. An hourglass network consisting of 3D convolutions, with three down-sampling blocks and three up-sampling blocks with residual connections, is leveraged for cost volume filtering. While 3D convolutions of kernel size $3 \times 3 \times 3$ are commonly used for relatively small disparity sizes [7, 22, 66], we observe it struggles with larger disparities when applied to high resolution images, especially since the disparity dimension is expected to model the probability distribution for the initial disparity prediction. However, it is impractical to naively increase the kernel size, due to the intensive memory consumption. In fact, even when setting kernel size to $5 \times 5 \times 5$ we observe unmanageable memory usage on an

80 GB GPU. This drastically limits the model's representation power when scaling up with large amount of training data. We thus develop "Axial-Planar Convolution" which decouples a single $3 \times 3 \times 3$ convolution into two separate convolutions: one over spatial dimensions (kernel size $K_s \times K_s \times 1$) and the other over disparity ($1 \times 1 \times K_d$), each followed by BatchNorm and ReLU. APC can be regarded as a 3D version of Separable Convolution [14] with the difference that we only separate the spatial and disparity dimensions without subdividing the channel into groups which sacrifices representation power. The disparity dimension is specially treated due to its uniquely encoded feature comparison within the cost volume. We use APC wherever possible in the hourglass network except for the downsampling and up-sampling layers.

Disparity Transformer (DT). While prior works [32, 63] introduced transformer architecture to unary feature extraction step to scale up stereo training, the cost filtering process is often overlooked, which remains an essential step in achieving accurate stereo matching by encapsulating correspondence information. Therefore, we introduce DT to further enhance the long-range context reasoning within the 4D cost volume. Given $\mathbf{V_C}$ obtained in Eq. (1), we first apply a 3D convolution of kernel size $4 \times 4 \times 4$ with stride 4 to downsize the cost volume. We then reshape the volume into a batch of token sequences, each with length of disparity. We apply position encoding before feeding it to a series (4 in our case) of transformer encoder blocks, where FlashAttention [16] is leveraged to perform multi-head self-attention [58]. The process can be written as:

$$\mathbf{Q_0} = \text{PE}\left(\mathbf{R}\left(\text{Conv}_{4\times4\times4}(\mathbf{V_C})\right)\right) \in \mathbb{R}^{(\frac{H}{16}\times\frac{W}{16})\times C\times\frac{D}{16}}$$

$$\text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = [\text{head}_1,\ldots,\text{head}_h]\mathbf{W}_O$$

$$\text{where head}_i = \text{FlashAttention}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i)$$

$$\mathbf{Q_1} = \text{Norm}\left(\text{MultiHead}(\mathbf{Q_0},\mathbf{Q_0},\mathbf{Q_0}) + \mathbf{Q_0}\right)$$

$$\mathbf{Q_2} = \text{Norm}\left(\text{FFN}(\mathbf{Q_1}) + \mathbf{Q_1}\right)$$

where $\mathbf{R}(\cdot)$ denotes reshape operation; $\text{PE}(\cdot)$ represents position encoding; $[\cdot,\cdot]$ denotes concatenation along the channel dimension; $\mathbf{W}_O$ is linear weights. The number of heads is $h = 4$ in our case. Finally, the DT output is up-sampled to the same size as $\mathbf{V_C}$ using trilinear interpolation and summed with hourglass output, as shown in Fig. 2.

Initial Disparity Prediction. We apply soft-argmin [27] to the filtered volume $\mathbf{V'_C}$ to produce an initial disparity:

$$d_0 = \sum_{d=0}^{\frac{D}{4}-1} d \cdot \text{Softmax}(\mathbf{V'_C})(d) \tag{2}$$

where d_0 is at 1/4 scale of the original image resolution.

3.3. Iterative Refinement

Given d_0, we perform iterative GRU updates to progressively refine disparity, which helps to avoid local optimum and accelerate convergence [66]. In general, the k-th update can be formulated as:

$$\mathbf{V}_{\text{corr}}(w',h,w) = \left\langle f_l^{(4)}(h,w), f_r^{(4)}(h,w')\right\rangle \tag{3}$$

$$\mathbf{F_V}(h,w) = [\mathbf{V'_C}(d_k,h,w), \mathbf{V}_{\text{corr}}(w-d_k,h,w)] \tag{4}$$

$$x_k = [\text{Conv}_v(\mathbf{F_V}), \text{Conv}_d(d_k), d_k, c] \tag{5}$$

$$z_k = \sigma\left(\text{Conv}_z([h_{k-1},x_k])\right) \tag{6}$$

$$r_k = \sigma\left(\text{Conv}_r([h_{k-1},x_k])\right) \tag{7}$$

$$\hat{h}_k = \tanh\left(\text{Conv}_h([r_k \odot h_{k-1}, x_k])\right) \tag{8}$$

$$h_k = (1-z_k)\odot h_{k-1} + z_k \odot \hat{h}_k \tag{9}$$

$$d_{k+1} = d_k + \text{Conv}_\Delta(h_k) \tag{10}$$

where $\odot$ denotes element-wise product; σ denotes sigmoid; $\mathbf{V}_{\text{corr}} \in \mathbb{R}^{\frac{W}{4}\times\frac{H}{4}\times\frac{W}{4}}$ is the pair-wise correlation volume; $\mathbf{F_V}$ represents the looked up volume features using latest disparity; $c = \text{ReLU}(f_c)$ encodes the context feature from left image, including STA adapted features (Sec. 3.1) which effectively guide the refinement process leveraging rich monocular priors.

We use three levels of GRU blocks to perform coarse-to-fine hidden state update in each iteration, where the initial hidden states are produced from context features $h_0^{(i)} = \tanh(f_c^{(i)}), i \in \{4,8,16\}$. At each level, attention-based selection mechanism [62] is leveraged to capture information at different frequencies. Finally, d_k is up-sampled to the full resolution using convex sampling [53].

3.4. Loss Function

The model is trained with the following objective:

$$\mathcal{L} = \left|d_0 - \overline{d}\right|_{\text{smooth}} + \sum_{k=1}^{K} \gamma^{K-k} \left\|d_k - \overline{d}\right\|_1 \tag{11}$$

where $\overline{d}$ represents ground-truth disparity; $|\cdot|_{\text{smooth}}$ denotes smooth L_1 loss; k is the iteration number; γ is set to 0.9, and we apply exponentially increasing weights [33] to supervise the iteratively refined disparity.

3.5. Synthetic Training Dataset

We created a large scale synthetic training dataset with NVIDIA Omniverse. This FoundationStereo Dataset (FSD) accounts for crucial stereo matching challenges such as reflections, low-texture surfaces, and severe occlusions. We perform domain randomization [54] to augment dataset diversity, including random stereo baseline, focal length, camera perspectives, lighting conditions and object configurations. Meanwhile, high-quality 3D assets with abundant textures and path-tracing rendering are leveraged to enhance realism in rendering and layouts. Fig. 4 displays some samples from our dataset including both structured indoor and outdoor scenarios, as well as more diversely randomized flying objects with various geometries and textures under complex yet realistic lighting. See the appendix for details.

Iterative Self-Curation. While synthetic data generation

Figure 4. **Left:** Samples from our FoundationStereo dataset (FSD), which consists of synthetic stereo images with structured indoor / outdoor scenes (top), as well as more randomized scenes with challenging flying objects and higher geometry and texture diversity (bottom). **Right:** The iterative self-curation process removes ambiguous samples inevitably produced from the domain randomized synthetic data generation process. Example ambiguities include severe texture repetition, ubiquitous reflections with limited surrounding context, and pure color under improper lighting.

in theory can produce unlimited amount of data and achieve large diversity through randomization, ambiguities can be inevitably introduced especially for less structured scenes with flying objects, which confuses the learning process. To eliminate those samples, we design an automatic iterative self-curation strategy. Fig. 4 demonstrates this process and detected ambiguous samples. We start with training an initial version of FoundationStereo on FSD, after which it is evaluated on FSD. Samples where BP-2 (Sec. 4.2) is larger than 60% are regarded as ambiguous samples and replaced by regenerating new ones. The training and curation processes are alternated to iteratively (twice in our case) update both FSD and FoundationStereo.

4. Experiments

4.1. Implementation Details

We implement FoundationStereo in PyTorch. The foundation model is trained on a mixed dataset consisting of our proposed FSD, together with Scene Flow [40], Sintel [4], CREStereo [31], FallingThings [56], InStereo2K [1] and Virtual KITTI 2 [5]. We train FoundationStereo using AdamW optimizer [36] for 200K steps with a total batch size of 128 evenly distributed over 32 NVIDIA A100 GPUs. The learning rate starts at 1e-4 and decays by 0.1 at 0.8 of the entire training process. Images are randomly cropped to 320×736 before feeding to the network. Data augmentations similar to [33] are performed. During training, 22 iterations are used in GRU updates. In the following, unless otherwise mentioned, we use the same foundation model for zero-shot inference using 32 refinement iterations.

4.2. Benchmark Datasets and Metric

Datasets. We consider five commonly used public datasets for evaluation: Scene Flow [40] is a synthetic dataset including three subsets: FlyingThings3D, Driving, and Monkaa. Middlebury [47] consists of indoor stereo image pairs with high-quality ground-truth disparity captured via structured light. ETH3D [48] provides grayscale stereo image pairs covering both indoor and outdoor scenarios. KITTI 2012 [18] and KITTI 2015 [41] datasets feature real-world driving scenes, where sparse ground-truth disparity maps are provided, which are derived from LIDAR sensors.

Methods	Middlebury BP-2	ETH3D BP-1	KITTI-12 D1	KITTI-15 D1
CREStereo++ [24]	14.8	4.4	4.7	5.2
DSMNet [77]	13.8	6.2	6.2	6.5
Mask-CFNet [45]	13.7	5.7	4.8	5.8
HVT-RAFT [8]	10.4	3.0	3.7	5.2
RAFT-Stereo [33]	9.4	3.3	4.7	5.5
Selective-IGEV [62]	9.2	5.7	4.5	5.6
IGEV [33]	8.8	4.0	5.2	5.7
Former-RAFT-DAM [79]	8.1	3.3	3.9	5.1
IGEV++ [67]	7.8	4.1	5.1	5.9
NMRF [20]	7.5	3.8	4.2	5.1
Ours (Scene Flow)	**5.5**	**1.8**	**3.2**	**4.9**
Selective-IGEV* [62]	7.5	3.4	3.2	4.5
Ours	**1.2**	**1.4**	**1.9**	**2.2**

Table 2. Zero-shot generalization results on four public datasets. The most commonly used metrics for each dataset were adopted. In the first block, all methods were trained only on Scene Flow. In the second block, methods are allowed to train on any existing datasets excluding the four target domains. The weights and parameters are fixed for evaluation.

Metrics. "EPE" computes average per-pixel disparity error. "BP-X" computes the percentage of pixels where the disparity error is larger than X pixels. "D1" computes the percentage of pixels whose disparity error is larger than 3 pixels and 5% of the ground-truth disparity.

4.3. Zero-Shot Generalization Comparison

Benchmark Evaluation. Tab. 2 exhibits quantitative comparison of zero-shot generalization results on four public real-world datasets. Even when trained solely on Scene Flow, our method outperforms the comparison methods consistently across all datasets, thanks to the efficacy of adapting rich monocular priors from vision foundation models. We further evaluate in a more realistic setup, allowing methods to train on any available dataset while excluding the target domain, to achieve optimal zero-shot inference results as required in practical applications.

In-the-Wild Generalization. We compare our foundation model against recent approaches that released their checkpoints trained on a mixture of datasets, to resemble the practical zero-shot application on in-the-wild images. Comparison methods include CroCo v2 [63], CREStereo [31], IGEV [66] and Selective-IGEV [62]. For each method, we select the best performing checkpoint from their public re-

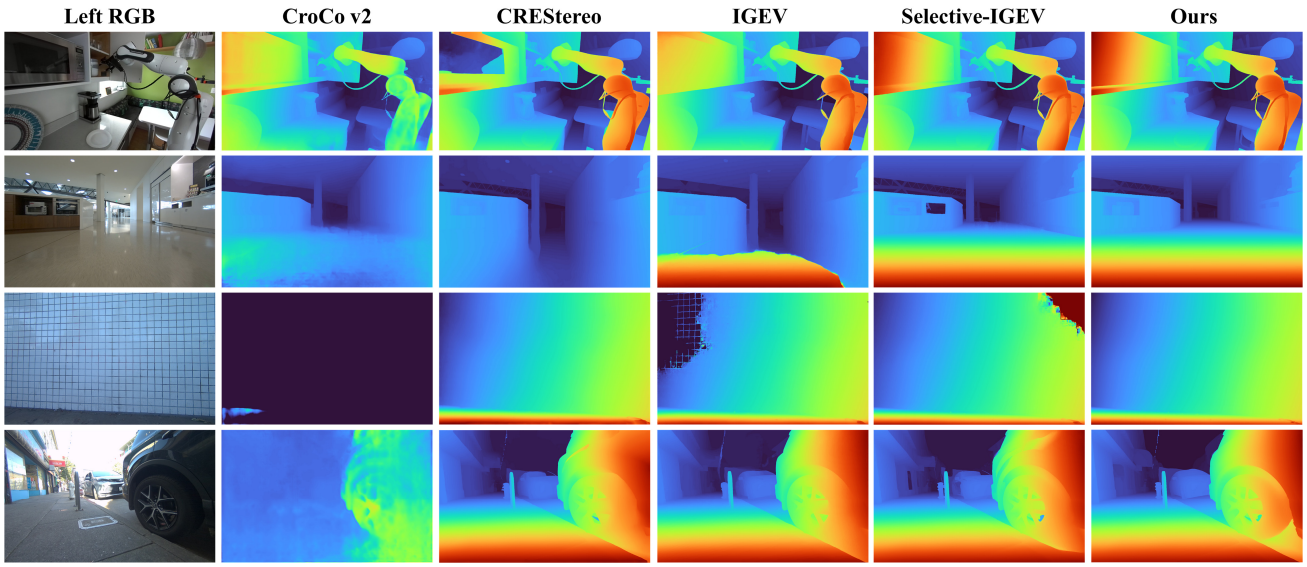

Figure 5. Qualitative comparison of zero-shot inference on in-the-wild images. For each comparison method we select the best performing checkpoint from their public release, which has been trained on a mixture of public datasets. These images exhibit challenging reflection, translucency, repetitive textures, complex illuminations and thin-structures, revealing the importance of our network architecture and large-scale training.

Method	LEAStereo [13]	GANet [76]	ACVNet [65]	IGEV-Stereo [66]	NMRF [20]	MoCha-Stereo [12]	Selective-IGEV [62]	Ours
EPE	0.78	0.84	0.48	0.47	0.45	0.41	0.44	**0.33**

Table 3. Comparison of methods trained / tested on the Scene Flow train / test sets, respectively.

Method	Zero-Shot	BP-0.5	BP-1.0	EPE
GMStereo [69]	✗	5.94	1.83	0.19
HITNet [52]	✗	7.83	2.79	0.20
EAI-Stereo [80]	✗	5.21	2.31	0.21
RAFT-Stereo [33]	✗	7.04	2.44	0.18
CREStereo [31]	✗	3.58	0.98	0.13
IGEV-Stereo [66]	✗	3.52	1.12	0.14
CroCo-Stereo [63]	✗	3.27	0.99	0.14
MoCha-Stereo [12]	✗	3.20	1.41	0.13
Selective-IGEV [62]	✗	3.06	1.23	0.12
Ours (finetuned)	✗	**1.26**	**0.26**	**0.09**
Ours	✓	2.31	1.52	0.13

Table 4. Results on ETH3D leaderboard (test set). All methods except for the last row have used ETH3D training set for fine-tuning. Our fine-tuned version ranks 1st on leaderboard at the time of submission. Last row is obtained via zero-shot inference from our foundation model.

lease. In this evaluation, the four real-world benchmark datasets [18, 41, 47, 48] have been used for training comparison methods, whereas they are not used in our fixed foundation model. Fig. 5 displays qualitative comparison on various scenarios, including a robot scene from DROID [28] dataset and custom captures covering indoor and outdoor.

4.4. In-Domain Comparison

Tab. 3 presents quantitative comparison on Scene Flow, where all methods are following the same officially divided train and test split. Our FoundationStereo model outperforms the comparison methods by a large margin, reducing the previous best EPE from 0.41 to 0.33. Although in-domain training is not the focus of this work, the results reflect the effectiveness of our model design.

Tab. 4 exhibits quantitative comparison on ETH3D leaderboard (test set). For our approach, we perform evaluations in two settings. First, we fine-tune our foundation model on a mixture of the default training dataset (Sec. 4.1) and ETH3D training set for another 50K steps, using the same learning rate schedule and data augmentation. Our model significantly surpasses the previous best approach by reducing more than half of the error rates and ranks 1st on leaderboard at the time of submission. This indicates great potential of transferring capability from our foundation model if in-domain fine-tuning is desired. Second, we also evaluated our foundation model without using any data from ETH3D. Remarkably, our foundation model's zero-shot inference achieves comparable or even better results than leading approaches that perform in-domain training.

4.5. Ablation Study

We investigate different design choices for our model and dataset. Unless otherwise mentioned, we train on a randomly subsampled version (100K) of FSD to make the experiment scale more affordable. Given Middlebury dataset's high quality ground-truth, results are evaluated on its training set to reflect zero-shot generalization. Since the focus of this work is to build a stereo matching foundation model with strong generalization, we do not deliberately limit model size while pursuing better performance.

STA Design Choices. As shown in Tab. 5, we first compare different vision foundation models for adapting rich monoc-

Row	Variations	BP-2
1	DINOv2-L [42]	2.46
2	DepthAnythingV2-S [74]	2.22
3	DepthAnythingV2-B [74]	2.11
4	DepthAnythingV2-L [74]	1.97
5	STA (a)	6.48
6	STA (b)	2.22
7	STA (c)	1.97
8	Unfreeze ViT	3.94
9	Freeze ViT	1.97

Table 5. Ablation study of STA module. Variations (a-c) correspond to Fig. 3. The choices adopted in our full model are highlighted in green.

Row	Variations	BP-2
1	RoPE	2.19
2	Cosine	1.97
3	1/32	2.06
4	1/16	1.97
5	Full	2.25
6	Disparity	1.97
7	Pre-hourglass	2.06
8	Post-hourglass	2.20
9	Parallel	1.97

Row	Variations	BP-2
10	(3,3,1), (1,1,5)	2.10
11	(3,3,1), (1,1,9)	2.06
12	(3,3,1), (1,1,13)	2.01
13	(3,3,1), (1,1,17)	1.97
14	(3,3,1), (1,1,21)	1.98
15	(7,7,1), (1,1,17)	1.99

Table 6. Ablation study of AHCF module. Left corresponds to DT, while right corresponds to APC. The choices adopted in our full model are highlighted in green.

ular priors, including different model sizes of DepthAnythingV2 [74] and DINOv2-Large [42]. While DINOv2 previously exhibited promising results in correspondence matching [17], it is not as effective as DepthAnythingV2 in the stereo matching task, possibly due to its less task-relevance and its limited resolution to reason high-precision pixel-level correspondence. We then study different design choices from Fig. 3. Surprisingly, while being simple, we found (c) significantly surpasses the alternatives. We hypothesize the latest feature before the final output head preserves high-resolution and fine-grained semantic and geometric priors that are suitable for subsequent cost volume construction and filtering process. We also experimented whether to freeze the adapted ViT model. As expected, unfreezing ViT corrupts the pretrained monocular priors, leading to degraded performance.

AHCF Design Choices. As shown in Tab. 6, for DT module we study different position embedding (row 1-2); different feature scale to perform transformer (row 3-4); transformer over the full cost-volume or only along the disparity dimension (row 5-6); different placements of DT module relative to the hourglass network (row 7-9). Specifically, RoPE [51] encodes relative distances between tokens instead of absolute positions, making it more adaptive to varying sequence lengths. However, it does not outperform cosine position embedding, probably due to the constant disparity size in 4D cost volume. While in theory, full volume attention provides larger receptive field, it is less effective than merely applying over the disparity dimension of the cost volume. We hypothesize the extremely large space of 4D cost volume makes it less tractable, whereas attention over disparity provides sufficient context for a better initial disparity prediction and subsequent volume feature lookup during GRU updates. Next, we compare different kernel sizes in APC (row 10-15), where the last dimension in each parenthesis corresponds to disparity dimension. We observe increasing benefits when enlarging disparity kernel size until it saturates at around 17.

Row	STA	AHCF		BP2
		APC	DT	
1				2.48
2	✓			2.21
3	✓	✓		2.16
4	✓		✓	2.05
5	✓	✓	✓	1.97

Row	FSD	BP2
1	✗	2.34
2	✓	1.15

Table 7. **Left:** Ablation study of proposed network modules. **Right:** Ablation study of whether to use FSD dataset when training the foundation model described in Sec. 4.1. The choices adopted in our full model are highlighted in green.

Effects of Proposed Modules. The quantitative effects are shown in Tab. 7 (left). STA leverages rich monocular priors which greatly enhances generalization to real images for ambiguous regions. DT and APC effectively aggregate cost volume features along spatial and disparity dimensions, leading to improved context for disparity initialization and subsequent volume feature look up during GRU updates. Fig. 3 further visualizes the resulting effects.

Effects of FoundationStereo Dataset. We study whether to include FSD dataset with the existing public datasets for training our foundation model described in Sec. 4.1. Results are shown in Tab. 7 (right).

5. Conclusion

We introduced FoundationStereo, a foundation model for stereo depth estimation that achieves strong zero-shot generalization across various domains without fine-tuning. We envision such a foundation model will facilitate broader adoption of stereo estimation models in practical applications. Despite its remarkable generalization, our method is not without limitations. First, our model is not yet optimized for efficiency, which takes 0.7s on image size of 375×1242 on NVIDIA A100 GPU. Future work could explore adapting distillation and pruning techniques applied to other vision foundation models [11, 82]. Second, our dataset FSD includes a limited collection of transparent objects. Robustness could be further enhanced by augmenting with a larger diversity of fully transparent objects during training.

References

[1] Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. InStereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63:1–11, 2020. 2, 6

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3

[3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 3

[4] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 611–625, 2012. 2, 3, 6

[5] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, 2020. 2, 6

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 3

[7] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018. 4

[8] Tianyu Chang, Xun Yang, Tianzhu Zhang, and Meng Wang. Domain generalized stereo matching via hierarchical visual transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9559–9568, 2023. 1, 2, 6

[9] Liyan Chen, Weihan Wang, and Philippos 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 (CVPR)*, pages 17235–17244, 2023. 1, 2

[10] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *ICLR*, 2023. 3

[11] Zigeng Chen, Gongfan Fang, Xinyin Ma, and Xinchao Wang. 0.1% data makes segment anything slim. *NeurIPS*, 2023. 8

[12] Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu 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 (CVPR)*, pages 27768–27777, 2024. 1, 2, 7

[13] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Proceedings of Neural Information Processing Systems (NeurIPS)*, 33:22158–22169, 2020. 7

[14] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258, 2017. 5

[15] 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. 1, 2

[16] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with io-awareness. *Proceedings of Neural Information Processing Systems (NeurIPS)*, 35:16344–16359, 2022. 5

[17] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21795–21806, 2024. 8

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 2, 6, 7

[19] Rui Gong, Weide Liu, Zaiwang Gu, Xulei Yang, and Jun Cheng. Learning intra-view and cross-view geometric knowledge for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20752–20762, 2024. 1, 2

[20] Tongfan Guan, Chen Wang, and Yun-Hui Liu. Neural Markov random field for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5459–5469, 2024. 6, 7

[21] Weiyu Guo, Zhaoshuo Li, Yongkui Yang, Zheng Wang, Russell H Taylor, Mathias Unberath, Alan Yuille, and Yingwei Li. Context-enhanced stereo transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 263–279, 2022. 2

[22] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3273–3282, 2019. 4

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[24] 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 International Conference on Computer Vision (ICCV)*, pages 3318–3327, 2023. 1, 2, 6

[25] 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

[26] 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)*, pages 9492–9502, 2024. 3

[27] 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 (ICCV)*, pages 66–75, 2017. 5

[28] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 7

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 1, 3

[30] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. *arXiv preprint arXiv:2406.09756*, 2024. 3

[31] 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 (CVPR)*, pages 16263–16272, 2022. 1, 2, 3, 6, 7

[32] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6197–6206, 2021. 1, 2, 5

[33] Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, pages 218–227, 2021. 1, 2, 3, 5, 6, 7

[34] 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. 1, 2

[35] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 3

[36] I Loshchilov. Decoupled weight decay regularization. *ICLR*, 2019. 6

[37] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. EdgeNeXt: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–20, 2022. 3

[38] Yamin Mao, Zhihua Liu, Weiming Li, Yuchao Dai, Qiang Wang, Yun-Tae Kim, and Hong-Seok Lee. UASNet: Uncertainty adaptive sampling network for deep stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6311–6319, 2021. 1, 2

[39] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283–287, 1976. 1

[40] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 2, 3, 6

[41] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 2, 6, 7

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 1, 3, 8

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 3

[44] 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 (PAMI)*, 2023. 2

[45] 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. 1, 2, 6

[46] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[47] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 31–42. Springer, 2014. 2, 6, 7, 1

[48] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and An-

dreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3260–3269, 2017. 2, 6, 7

[49] 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 (CVPR)*, pages 13906–13915, 2021. 1, 2

[50] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 280–297, 2022. 1, 2

[51] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 8

[52] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. HITNet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14362–14372, 2021. 7

[53] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2020. 2, 5

[54] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017. 5

[55] 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 (PAMI)*, 2024. 1, 2

[56] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2038–2041, 2018. 2, 3, 6

[57] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *Conference on Robot Learning (CoRL)*, pages 306–316, 2018. 1

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 5

[59] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. IRS: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2021. 2, 3

[60] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 3

[61] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual slam. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916, 2020. 2, 3

[62] 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 (CVPR)*, pages 19701–19710, 2024. 1, 2, 5, 6, 7

[63] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 17969–17980, 2023. 1, 2, 5, 6, 7

[64] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17868–17879, 2024. 3

[65] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12981–12990, 2022. 7

[66] 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, 5, 6, 7

[67] Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Junda Cheng, Chunyuan Liao, and Xin Yang. IGEV++: Iterative multi-range geometry encoding volumes for stereo matching. *arXiv preprint arXiv:2409.00638*, 2024. 2, 6

[68] Haofei Xu and Juyong Zhang. AANet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1959–1968, 2020. 1, 2

[69] 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 (PAMI)*, 2023. 7

[70] 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 (CVPR)*, pages 5515–5524, 2019. 2

[71] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. DrivingStereo: A large-scale

dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 899–908, 2019. 2

[72] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 3

[73] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381, 2024. 1, 3

[74] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2, 3, 4, 8

[75] Menglong Yang, Fangrui Wu, and Wei Li. WaveletStereo: Learning wavelet coefficients of disparity map in stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12885–12894, 2020. 1, 2

[76] 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 (CVPR)*, pages 185–194, 2019. 7

[77] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 420–439, 2020. 1, 2, 6

[78] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 698–714, 2020. 3

[79] Yongjian Zhang, Longguang Wang, Kunhong Li, Yun Wang, and Yulan Guo. Learning representations from foundation models for domain generalized stereo matching. In *European Conference on Computer Vision*, pages 146–162. Springer, 2024. 1, 2, 6

[80] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Yong Zhao, Yitong Yang, and Ting Ouyang. EAI-Stereo: Error aware iterative network for stereo matching. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 315–332, 2022. 7

[81] 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 (CVPR)*, pages 1327–1336, 2023. 1, 2

[82] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 8

FoundationStereo: Zero-Shot Stereo Matching

Supplementary Material

5.1. ETH3D Leaderboard

At the time of submission, our fine-tuned model ranks 1st on the ETH3D leaderboard, significantly outperforming both published and unpublished works. The screenshot is shown in Fig. 6.

5.2. More Ablation Study on Synthetic Data

We study the effectiveness of self-curation pipeline introduced in Sec. 3.5. When disabling the self-curation while keeping the same data size, the synthetic dataset involves ambiguous samples that confuse the learning process, leading to slight performance drop when evaluated on Middlebury [47] dataset.

Variation	BP2
W/ self-curation	1.15
W/o self-curation	1.27

Table 8. Effectiveness of self-curation pipeline when generating synthetic data.

5.3. More Details of Synthetic Data Generation

Tooling and Assets. The dataset generation is built on NVIDIA Omniverse. We use RTX path-tracing with 32 to 128 samples per pixel for high-fidelity photorealistic rendering. The data generation is performed across 48 NVIDIA A40 GPUs for 10 days. There are more than 5K object assets collected from varying sources including artist designs and 3D scanning with high-frequency geometry details. Object assets are divided into the groups of: furniture, open containers, vehicles, robots, floor tape, free-standing walls, stairs, plants, forklifts, dynamically animated digital humans, other obstacles and distractors. Each group is defined with a separate randomization range for sampling locations, scales and appearances. In addition, we curated 12 large scene models (Fig. 7), 16 skybox images, more than 150 materials, and 400 textures for tiled wrapping on object geometries for appearance augmentation. These textures are obtained from real-world photos and procedurally generated random patterns.

Camera Configuration. For each data sample, we first randomly sample the stereo baseline camera focal length to diversify the coverage of field-of-views and disparity distributions. Next, objects are spawned into the scene in two different methods to randomize the scene configuration: 1) camera is spawned in a random pose, and objects are added relative to the camera at random locations; 2) objects are spawned near a random location, and the camera is spawned nearby and oriented to the center of mass of the object clutter.

Layout Configuration. We generate layouts in two kinds of styles: chaotic and realistic. Such combination of the more realistic structured layouts with the more randomized setups with flying objects has been shown to benefit sim-to-real generalization [57]. Specifically, chaotic-style scenes involve large number of flying distractors and simple scene layouts which consists of infinitely far skybox and a background plane. The lighting and object appearances (texture and material) are highly randomized. The realistic-style data uses indoor and outdoor scene models where the camera is restricted to locate at predefined areas. Object assets are dropped and applied with physical properties for collision. The simulation is performed randomly between 0.25 to 2 seconds to create physically realistic layouts with no penetration, involving both settled and falling objects. Materials and scales native to object assets are maintained and more natural lighting is applied. Among the realistic-style data, we further divide the scenes into three types which determine what categories of objects are selected to compose the scene for more consistent semantics:

- Navigation - camera poses are often in parallel to the ground and objects are often spawned further away. Objects such as free-standing walls, furniture, and digital humans are sampled with higher probability.
- Driving - camera is often in parallel to the ground above the ground and objects are often spawned further away. Objects such as vehicles, digital humans, poles, signs and speed bumps are sampled with higher probability.
- Manipulation - camera is oriented to face front or downward as in ego-centric views and objects are often spawned in closer range to resemble interaction scenarios. Objects such as household or grocery items, open containers, robotic arms are sampled with higher probability.

Lighting Configuration. Light types include global illumination, directed sky rays, lights baked-into 3D scanned assets, and light spheres which add dynamic lighting when spawned near to surfaces. Light colors, intensities and directions are randomized. Lighting vibes such as daytime, dusk and night are included within the random sampling ranges.

Disparity Distribution. Fig. 8 shows the disparity distribution of our FSD dataset.

5.4. Acknowledgement

We would like to thank Gordon Grigor, Jack Zhang, Karsten Patzwaldt, Hammad Mazhar and other NVIDIA Isaac team

ETH3D

Home | Stereo | SLAM | About | Submit

Method	Info	all	lakes. 1l	lakes. 1s	sand box 1l	sand box 1s	stora. room 1l	stora. room 1s	stora. room 2l	stora. room 2s	stora. room 2 1l	stora. room 2 1s	stora. room 2 2l	stora. room 2 2s	stora. room 3l	stora. room 3s	tunnel 1l	tunnel 1s	tunnel 2l
FoundationStereo		**0.26** 1	0.29 30	1.25 8	0.24 64	0.10 59	0.41 2	0.07 36	**0.57** 1	0.80 4	0.24 9	0.03 2	0.09 3	0.37 48	**0.46** 1	**0.12** 1	0.03 314	0.03 252	**0.00** 1
MonSter		0.46 2	0.23 14	1.44 13	0.05 10	0.77 268	1.17 10	0.01 2	3.17 23	**0.72** 1	0.28 14	0.18 10	**0.07** 1	**0.04** 1	0.83 9	0.17 7	**0.00** 1	**0.00** 1	**0.00** 1
dual_stereo		0.56 3	0.32 51	1.08 3	0.07 16	0.11 68	0.77 6	0.03 15	3.60 43	1.17 14	**0.12** 1	0.03 2	0.20 14	0.11 4	2.77 83	0.51 69	**0.00** 1	0.14 385	**0.00** 1
RAStereo		0.68 4	0.42 138	3.15 189	0.18 55	0.52 235	1.17 10	0.02 11	3.16 22	0.98 9	0.51 56	1.43 48	0.16 7	0.22 13	0.75 6	0.83 125	**0.00** 1	**0.00** 1	**0.00** 1
GIP-stereo		0.70 5	0.44 155	1.70 34	1.52 250	0.75 264	1.48 33	0.05 25	4.18 72	1.10 10	0.18 5	0.42 18	0.58 75	0.38 51	0.65 3	0.46 54	**0.00** 1	**0.00** 1	**0.00** 1
DEFOM-Stereo		0.70 5	0.29 30	1.62 24	0.16 49	0.06 24	1.95 75	0.45 160	0.88 2	0.74 2	0.55 58	0.51 20	0.14 6	0.09 3	6.32 191	0.20 11	**0.00** 1	**0.00** 1	**0.00** 1
GREAT-IGEV		0.72 7	0.29 30	1.91 45	1.31 221	0.40 199	1.41 30	0.08 41	2.29 7	2.48 40	0.18 5	0.69 24	0.10 4	0.34 43	1.72 50	0.37 32	**0.00** 1	0.09 345	**0.00** 1
IGEV-Stereo++		0.74 8	0.23 14	1.31 10	0.06 13	1.62 329	1.95 75	0.01 2	3.29 27	1.81 20	0.38 32	1.68 58	0.10 4	1.16 136	0.91 11	0.19 10	0.01 203	0.01 153	**0.00** 1
GLC_STEREO		0.75 9	0.29 30	1.19 6	0.55 117	**0.01** 1	1.98 79	0.57 197	3.07 16	3.02 83	0.90 72	1.48 49	0.19 10	0.40 55	0.93 12	0.36 28	**0.00** 1	0.01 153	**0.00** 1
rvit_stereo_0081_aga		0.76	0.51	3.29	0.11	0.25	1.73	0.22	3.68	2.54	0.26	0.16	0.27	0.30	1.31	0.36	**0.00**	**0.00**	**0.00**

Figure 6. ETH3D leaderboard screenshot. Our fine-tuned foundation model (red box) ranks 1st at the time of submission.

Figure 7. Examples scene models involving factory, hospital, wood attic, office, grocery store and warehouse. In the third column, we demonstrate an example of metallic material randomization being applied to augment scene diversity. The last column shows comparison of a warehouse between the real and our simulated digital twin in high fidelity.

members for their tremendous engineering support and valuable discussions.

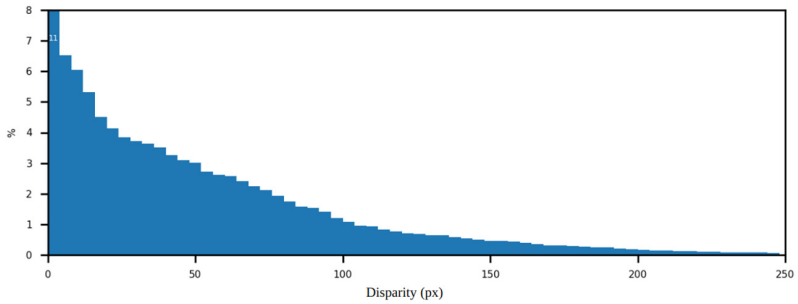

Figure 8. Disparity distribution in our proposed FSD.