# FoundationStereo: Zero-Shot Stereo Matching

## Supplementary Material

## 6. 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of submission, our fine-tuned model ranks 1st on the ETH3D leaderboard, significantly outperforming both published and unpublished works.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.

## 7. Middlebury Leaderboard

At the time of submission, our fine-tuned model ranks 1st on the Middlebury leaderboard, significantly outperforming both published and unpublished works. The screenshot is shown in Fig. 8.

## 8. More Ablation Study on Synthetic Data

**Effects of Self-Curation.** 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 [51] 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.

**Effects of FSD for Other Methods.** We train representative works IGEV and Selective-IGEV on FSD only. As shown in the table below, for both methods, our proposed FSD effectively boosts the performance compared to the commonly used Scene Flow dataset.

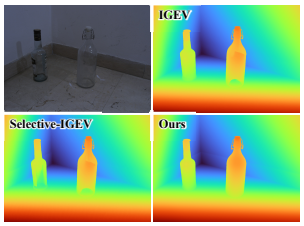
| Methods | Train data | Middlebury BP-2 | ETH3D BP-1 | KITTI-12 D1 | KITTI-15 D1 |
|---|---|---|---|---|---|
| IGEV | Scene Flow | 8.8 | 4.0 | 5.2 | 5.7 |
| IGEV | FSD | 7.8 | 3.5 | 3.2 | 4.7 |
| Selective-IGEV | Scene Flow | 9.2 | 5.7 | 4.5 | 5.6 |
| Selective-IGEV | FSD | 7.9 | 3.5 | 3.0 | 4.4 |

Table 9. Effects of FSD for other methods.

## 9. Results on Translucent Objects

We evaluate on Booster [48] (half resolution), which is a challenging dataset consisting of specular and transparent objects. We compare with the most competitive methods from Fig. 5 (main paper) in the zero-shot setting. The quantitative and qualitative results are shown below.

| Methods | Half | | | |
|---|---|---|---|---|
| | BP1 | BP-2 | BP-3 | EPE |
| Selective-IGEV | 23.8 | 15.0 | 12.0 | 6.6 |
| IGEV | 30.8 | 22.3 | 19.0 | 22.7 |
| Ours | **19.0** | **9.6** | **6.7** | **2.2** |




## 10. More Results on Middlebury Dataset

We compare with competitive methods that released their public weights in zero-shot on Middlebury, shown in below table. Since NMRF [22] did not report their evaluated Middlebury resolution, we rerun their released weights on all resolutions. At full resolution, maximum disparity 320 is used for FoundationStereo. Across all resolutions, ours significantly outperforms baselines. We also report the peak memory usage and running time averaged across the dataset on the same hardware, particularly single GPU 3090. On half and quarter resolutions, our peak memory occurs at STA module. On full resolution, it occurs at DT module. Despite the speed limitation which is not the focus when developing this work, ours can successfully run on a desktop GPU. Pruning or distillation remains an interesting future work to improve speed and memory footprint.

| Methods | Full | | | Half | | | Quarter | | |
|---|---|---|---|---|---|---|---|---|---|
| | BP-2 | peak mem (G) | time (s) | BP-2 | peak mem (G) | time (s) | BP-2 | peak mem (G) | time (s) |
| Selective-IGEV[61] | 12.9 | 6.9 | 2.52 | 9.2 | 1.7 | 0.72 | 7.0 | 0.5 | 0.25 |
| IGEV[33] | 13.1 | 6.3 | 2.06 | 8.8 | 1.6 | 0.53 | 6.4 | 0.5 | 0.18 |
| IGEV++[66] | 12.7 | 13.1 | 2.12 | 7.8 | 3.4 | 0.50 | 6.3 | 0.9 | 0.15 |
| NMRF[20] | 35.3 | 8.1 | 0.95 | 10.9 | 1.8 | 0.20 | 5.0 | 0.5 | 0.05 |
| Ours | **4.8** | 18.5 | 8.14 | **1.1** | 10.5 | 2.97 | **1.3** | 2.3 | 0.55 |

Table 10. Results on varying resolutions in Middlebury.

## 11. 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

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 (bottom) and our simulated digital twin (top) in high fidelity.

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 [62]. Specifically, chaotic-style scenes involve large number of flying distractors and simple scene

| Date | Name | Res | Weight Avg | Austr MP: 5.6 nd: 290 im0 im1 GT nonocc | AustrP MP: 5.6 nd: 290 im0 im1 GT nonocc | Bicyc2 MP: 5.6 nd: 250 im0 im1 GT nonocc | Class MP: 5.7 nd: 610 im0 im1 GT nonocc | ClassE MP: 5.7 nd: 610 im0 im1 GT nonocc | Compu MP: 1.5 nd: 256 im0 im1 GT nonocc | Crusa MP: 5.5 nd: 800 im0 im1 GT nonocc | CrusaP MP: 5.5 nd: 800 im0 im1 GT nonocc | Djemb MP: 5.7 nd: 320 im0 im1 GT nonocc | DjembL MP: 5.7 nd: 320 im0 im1 GT nonocc | Hoops MP: 5.7 nd: 410 im0 im1 GT nonocc | Livgrm MP: 5.9 nd: 320 im0 im1 GT nonocc | Nkuba MP: 5.5 nd: 570 im0 im1 GT nonocc | Plants MP: 5.6 nd: 320 im0 im1 GT nonocc | Stairs MP: 5.2 nd: 450 im0 im1 GT nonocc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 02/03/25 | FoundationStereo | F | **1.84** 1 | 2.46 3 | 1.71 2 | **1.36** 1 | 0.79 2 | 5.19 20 | **0.53** 1 | **0.93** 1 | **0.84** 1 | 0.93 7 | **2.41** 1 | **3.39** 1 | 3.45 10 | **3.28** 1 | 1.82 2 | **1.17** 1 |
| 08/07/24 | AIO-Stereo | F | 2.36 2 | 2.38 2 | 1.71 2 | 3.22 32 | 0.85 4 | 5.83 29 | 1.24 10 | 1.42 10 | 1.32 10 | 1.03 15 | 4.49 15 | 4.81 9 | 2.43 4 | 3.61 3 | 2.12 6 | 3.63 11 |
| 11/03/24 | DEFOM-Stereo | F | 2.39 3 | 2.82 10 | 2.21 11 | 1.53 2 | 1.01 9 | 5.24 22 | 0.88 3 | 1.40 9 | 1.14 4 | 0.85 4 | 2.64 5 | 9.10 34 | **2.18** 1 | 5.50 8 | 2.49 14 | 1.67 2 |
| 08/01/24 | RSM | F | 2.40 4 | 2.66 7 | 1.88 7 | 3.18 30 | 0.91 6 | 5.80 28 | 1.34 14 | 1.35 5 | 1.16 5 | 0.93 7 | 3.35 8 | 3.96 4 | 2.88 6 | 4.38 6 | 2.01 4 | 4.15 13 |
| 11/13/23 | Selective-IGEV | F | 2.51 5 | 2.54 6 | 1.86 6 | 2.51 19 | 1.12 13 | 7.22 43 | 1.23 9 | 1.36 6 | 1.17 6 | 1.16 20 | 4.48 14 | 4.83 10 | 2.99 7 | 3.79 5 | 2.26 10 | 4.72 18 |
| 12/16/24 | RPS | F | 2.61 6 | 2.46 3 | 1.71 2 | 3.92 41 | 0.79 2 | 5.19 20 | 2.44 33 | **0.93** 1 | **0.84** 1 | 0.93 7 | **2.41** 1 | **3.39** 1 | 3.45 10 | **3.28** 1 | 1.82 2 | 11.6 87 |
| 10/30/24 | MonoStereo | F | 2.64 7 | 3.72 30 | **1.68** 1 | 1.77 6 | 1.05 10 | 10.5 75 | 0.88 3 | 1.27 4 | 0.97 3 | **0.63** 1 | 4.39 13 | 8.10 24 | 4.59 24 | 3.70 4 | **1.73** 1 | 2.69 5 |
| 08/12/24 | PointerNet | F | 2.69 8 | 2.67 8 | 1.84 5 | 3.21 31 | 1.51 17 | 7.52 48 | 1.29 11 | 1.54 12 | 1.17 6 | 1.09 18 | 3.59 9 | 3.96 4 | 3.10 8 | 5.60 9 | 2.29 12 | 4.27 14 |
| 02/10/25 | GREAT-IGEV | F | 2.81 9 | 3.30 16 | 2.44 15 | 2.31 15 | 0.96 7 | 7.12 42 | 1.17 7 | 1.38 7 | 1.36 11 | 1.04 16 | 3.89 11 | 3.82 3 | 4.66 25 | 6.24 14 | 2.17 7 | 4.65 16 |
| 11/05/24 | coffe_stereo | F | 2.82 10 | 2.70 9 | 1.98 9 | 1.87 10 | **0.61** 1 | 3.32 7 | 2.45 34 | 1.07 3 | 1.30 9 | 1.02 13 | 2.63 4 | 4.13 6 | **2.18** 1 | 8.38 22 | 2.27 11 | 11.6 87 |
| 11/10/22 | DLNR | F | 3.20 11 | 2.91 12 | 2.37 14 | 2.18 13 | 1.67 20 | 3.21 5 | 1.37 15 | 1.66 13 | 1.66 15 | 1.11 19 | 6.25 28 | 7.07 20 | 3.45 10 | 8.90 27 | 4.43 32 | 2.91 7 |
| 06/14/24 | IGEV++ | F | 3.23 12 | 3.24 15 | 2.46 16 | 4.12 49 | 1.15 14 | 6.71 39 | 1.38 16 | 1.53 11 | 1.52 12 | 1.02 13 | 4.57 16 | 4.68 8 | 5.41 34 | 7.68 20 | 2.22 8 | 4.68 17 |
| 11/28/24 | DEFOM-Stereo_RVC | F | 3.28 13 | 3.50 20 | 2.61 19 | 2.41 17 | 0.87 5 | 2.51 2 | 0.89 5 | 1.38 7 | 1.26 8 | 0.97 12 | 6.35 29 | 10.8 50 | 2.43 4 | 11.0 38 | 3.03 18 | 5.00 22 |
| 06/27/24 | CAS++ | F | 3.33 14 | 4.27 44 | 3.72 53 | 3.17 29 | 2.17 32 | **2.44** 1 | 1.33 13 | 2.24 23 | 2.01 20 | 1.47 32 | 4.04 12 | 8.15 25 | 4.97 30 | 5.80 11 | 3.73 29 | 3.04 8 |
| 02/21/24 | ClearDepth | F | 3.48 15 | 4.14 41 | 3.16 37 | 2.81 23 | 1.95 24 | 4.55 14 | 2.36 31 | 1.73 16 | 1.70 18 | 1.25 27 | 5.46 22 | 11.2 57 | 3.12 9 | 7.30 18 | 3.70 28 | 3.45 10 |
| 06/06/24 | MoCha-V2 | F | 3.51 16 | 2.52 5 | 1.95 8 | 2.25 14 | 1.47 16 | 4.61 15 | 0.98 6 | 7.35 78 | 8.07 88 | 0.66 2 | 2.95 6 | 4.18 7 | 4.46 21 | 5.70 10 | 2.54 15 | 2.70 6 |
| 06/05/24 | MGS-Stereo | F | 3.57 17 | 3.62 23 | 2.93 26 | 3.43 35 | 2.66 43 | 6.24 36 | 2.54 36 | 2.04 20 | 2.15 24 | 1.23 26 | 5.81 25 | 8.40 28 | 3.56 14 | 6.48 15 | 3.18 21 | 4.81 19 |
| 10/28/23 | SAMTormer | F | 3.63 18 | 3.84 35 | 2.95 28 | 1.85 9 | 0.98 8 | 3.31 6 | 2.02 27 | 1.71 14 | 1.67 16 | 0.82 3 | 5.50 23 | 10.4 46 | 4.57 23 | 11.7 46 | 3.92 30 | 3.41 9 |
| 06/13/22 | EAI-Stereo | F | 3.68 19 | 4.02 37 | 3.32 40 | 2.48 18 | 1.42 15 | 4.19 12 | 2.37 32 | 2.18 22 | 2.01 20 | 1.16 20 | 10.2 48 | 8.84 32 | 4.00 17 | 7.15 17 | 3.14 20 | 6.44 31 |
| 11/10/21 | CREStereo | F | 3.71 20 | 4.73 51 | 3.94 56 | 5.07 69 | 1.96 25 | 3.02 4 | 1.42 17 | 2.28 24 | 2.05 23 | 1.51 34 | 6.86 30 | 6.35 13 | 4.25 19 | 6.01 13 | 4.60 35 | 5.49 26 |
| 09/08/24 | RSD | F | 3.73 21 | **2.13** 1 | 1.98 9 | 1.71 3 | 2.03 28 | 2.63 3 | 0.87 2 | 8.66 86 | 9.69 105 | 0.96 11 | 2.54 3 | 6.82 17 | 2.34 3 | 7.76 21 | 2.23 9 | 2.57 4 |
| 02/28/24 | AKD_Stereo | F | 3.87 22 | 4.21 43 | 3.53 48 | 3.91 40 | 1.08 11 | 7.63 50 | 4.75 61 | 1.72 15 | 1.60 13 | 1.18 23 | 5.26 21 | 9.62 37 | 3.66 15 | 7.63 19 | 3.23 23 | 5.37 25 |
| 03/04/24 | ET_Stereo | F | 4.00 23 | 4.38 45 | 3.33 41 | 2.85 24 | 1.53 19 | 7.84 53 | 2.61 37 | 1.91 18 | 1.82 19 | 1.05 17 | 5.08 20 | 8.72 31 | 7.52 50 | 8.81 26 | 3.09 19 | 4.82 20 |
| 11/01/24 | GIP-stereo | F | 4.03 24 | 2.86 11 | 2.31 13 | 2.64 21 | 1.76 21 | 5.00 16 | 1.55 18 | 6.75 73 | 7.30 80 | 0.86 6 | 4.63 17 | 7.24 21 | 3.51 13 | 10.6 36 | 2.06 5 | 2.37 3 |
| 10/09/23 | EGLCR-Stereo | F | 4.03 25 | 4.69 48 | 2.46 16 | 3.70 37 | 2.99 55 | 10.7 80 | 2.48 35 | 1.95 19 | 1.63 14 | 0.94 10 | 5.76 24 | 8.17 26 | 3.84 16 | 10.3 35 | 2.99 17 | 4.87 21 |
| 03/04/24 | AEACV | F | 4.15 26 | 5.53 69 | 2.98 30 | 2.54 20 | 3.23 59 | 3.42 9 | 1.57 19 | 2.85 28 | 2.99 33 | 1.22 24 | 4.63 17 | 5.96 12 | 4.36 20 | 12.9 59 | 5.41 53 | 4.08 12 |
| 10/30/23 | LoS | F | 4.20 27 | 5.85 75 | 4.92 103 | 4.64 60 | 2.77 48 | 3.92 10 | 1.32 12 | 2.36 25 | 2.17 25 | 1.81 40 | 8.18 37 | 6.58 14 | 4.55 22 | 8.57 24 | 4.57 33 | 5.06 24 |

Figure 8. Middlebury leaderboard screenshot. Our fine-tuned foundation model (red box) ranks 1st at the time of submission.

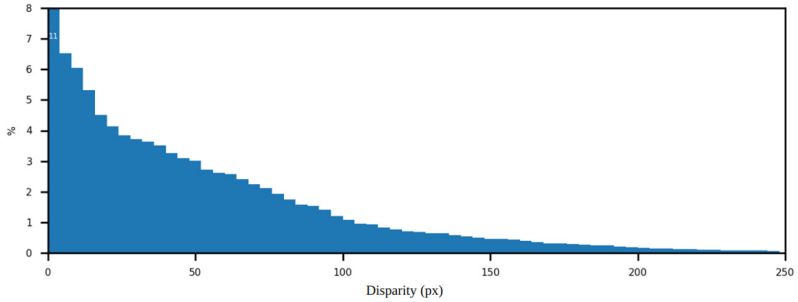



Figure 9. Disparity distribution in our proposed FSD.

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. 9 shows the disparity distribution of our FSD dataset.

## 12. Acknowledgement

We would like to thank Gordon Grigor, Jack Zhang, Xutong Ren, Karsten Patzwaldt, Hammad Mazhar and other NVIDIA Isaac team members for their tremendous engineering support and valuable discussions.