Project Page

Boosting Omnidirectional Stereo Matching with a Pre-trained Depth Foundation Model

Jannik Endres[1,2] Oliver Hahn[2] Charles Corbière[1]
Simone Schaub-Meyer[2,3] Stefan Roth[2,3] Alexandre Alahi[1]

[1] EPFL [2] Technische Universität Darmstadt [3] hessian.AI

TL ; DR

Given a pair of equirectangular images captured by two vertically stacked omnidirectional cameras, DFI-OmniStereo (*Depth Foundation Model-based Iterative Omnidirectional Stereo Matching*) integrates a large-scale pre-trained monocular relative depth foundation model into an iterative stereo matching approach. This method improves depth estimation accuracy, significantly outperforming the previous state-of-the-art method on the Helvipad dataset.

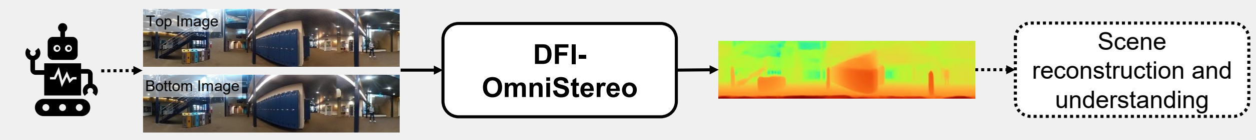

Figure 1. **DFI-OmniStereo** requires two omnidirectional images in a top-bottom camera configuration and predicts a dense omnidirectional disparity map, which can be converted into a depth map. These depth predictions can be used for subsequent scene reconstruction and scene understanding tasks.

Motivation

- **Omnidirectional depth** information crucial for **many robotics applications**
- **Conventional approach:** Depth measurement with **LiDAR** sensor is expensive and sparse
- **Alternative approach:** Depth estimation using **omnidirectional stereo matching**
- **Challenge: Low accuracy** of existing omnidirectional stereo matching methods due to **limited** real-world **data**

⇒ **Idea:** Leverage a **pre-trained foundation model** for **relative depth** to improve generalization and training sample efficiency in **omnidirectional stereo matching**

Method

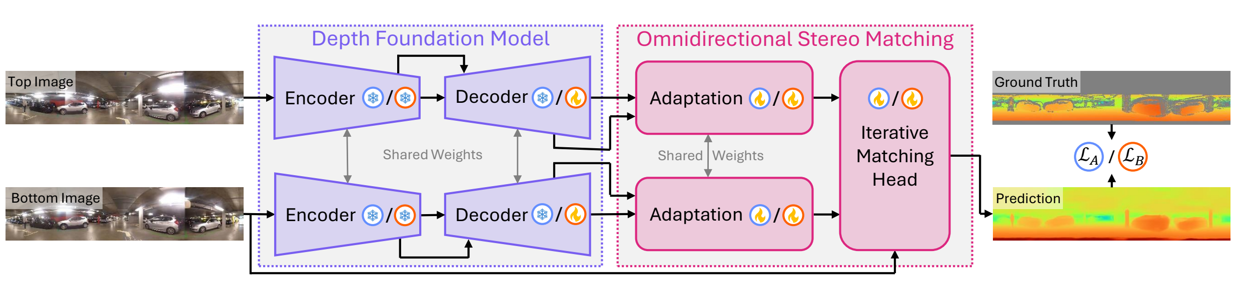

Figure 2. **DFI-OmniStereo's architecture** comprises a depth foundation model (purple) for feature extraction and an omnidirectional stereo matching head (pink). The training consists of two stages: stage A (blue) and stage B (orange).

DFI-OmniStereo Architecture

- **Monocular depth foundation model** (Depth Anything V2 [1]) to extract a **relative depth map** and **feature maps**
- Good initial representation for stereo matching due to **strong task relationship** between **relative depth** and **disparity estimation**
- **Omnidirectional stereo matching** head inspired by an **iterative optimization-based stereo matching architecture** (IGEV-Stereo [2])
- **Adaptation module: Bilinear interpolation** and a learnable **linear projection** to adjust feature dimensions

Training Strategy

Stage A – Adapt the stereo matching head to the foundation model's **feature representation**, the **camera setup**, and the **omnidirectional imagery**:

- **Foundation model frozen**, stereo matching and adaptation modules trainable
- **L1-based loss** $\mathcal{L}_A(\{\hat{\boldsymbol{d}}_i\}_{i=0}^N) = \mathcal{L}_{sL_1}\left(\hat{\boldsymbol{d}}_0, \boldsymbol{d}\right) + \sum_{i=1}^N \gamma^{N-i} \mathcal{L}_{L_1}\left(\hat{\boldsymbol{d}}_i, \boldsymbol{d}\right)$ with $\hat{\boldsymbol{d}}_i$: predicted disparities at iteration i, N: total iterations, $\boldsymbol{d}$: ground-truth disparity, $\mathcal{L}_{L_1}$: **L1 loss** and $\mathcal{L}_{sL_1}$: **smooth L1 loss**

Stage B – Scale-invariant fine-tuning of the **foundation model** to the **omnidirectional imagery** and the task of **stereo matching**:

- **Foundation model encoder frozen**, remaining modules trainable
- **SILog loss** $\mathcal{L}_B\left(\{\hat{\boldsymbol{d}}_i\}_{i=0}^N\right) = \mathcal{L}_{SIL}\left(\hat{\boldsymbol{d}}_0, \boldsymbol{d}\right) + \sum_{i=1}^N \gamma^{N-i} \mathcal{L}_{SIL}\left(\hat{\boldsymbol{d}}_i, \boldsymbol{d}\right)$, where $\mathcal{L}_{SIL}(\hat{\boldsymbol{d}}, \boldsymbol{d}) = \frac{1}{n}\sum_{j=1}^n \delta_{\log}\left(\hat{d}_j, d_j\right)^2 - \frac{\lambda}{n^2}\left(\sum_{j=1}^n \delta_{\log}\left(\hat{d}_j, d_j\right)\right)^2$ with $\delta_{log}\left(\hat{d}, d\right) = \log \hat{d} - \log d$

Results

Table 1. **Comparative results of omnidirectional stereo depth estimation on the Helvipad [3] test split.**

Method	Stereo Setting	Disparity (°)				Depth (m)			
		MAE ↓	RMSE ↓	MARE ↓	LRCE ↓	MAE ↓	RMSE ↓	MARE ↓	LRCE ↓
PSMNet [4]	Conventional	0.286	0.496	0.248	–	2.509	5.673	0.176	1.809
360SD-Net [5]	Omnidirectional	0.224	0.419	0.191	–	2.122	5.077	0.152	0.904
IGEV-Stereo [2]	Conventional	0.225	0.423	0.172	–	1.860	4.474	0.146	1.203
360-IGEV-Stereo [3]	Omnidirectional	0.188	0.404	0.146	**0.054**	1.720	4.297	0.130	**0.388**
DFI-OmniStereo (Ours)	Omnidirectional	**0.158**	**0.338**	**0.120**	0.058	**1.463**	**3.767**	**0.108**	0.397

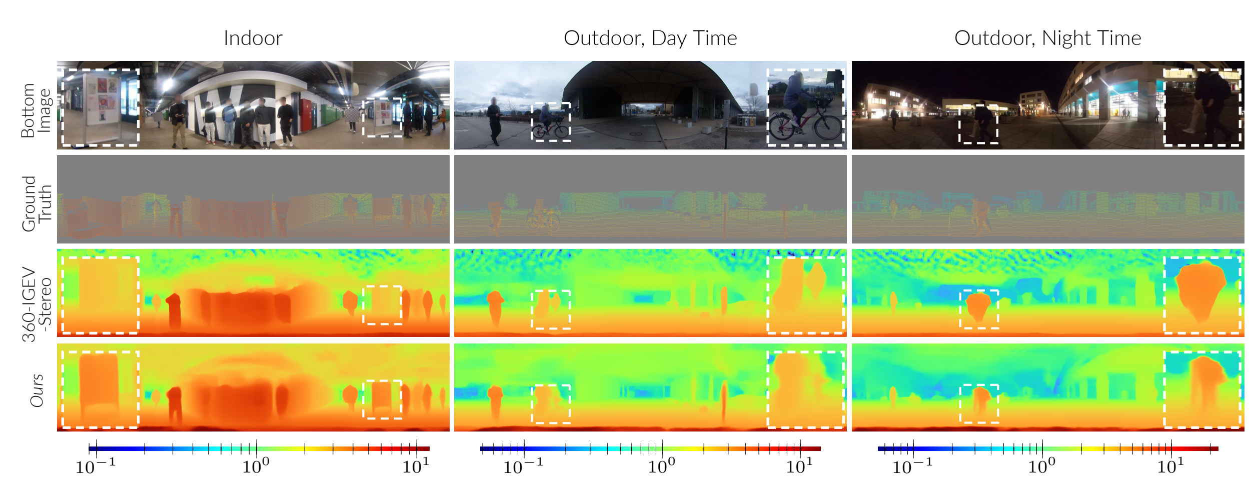

Figure 3. **Qualitative comparison of disparity map predictions (°) on the Helvipad [3] test split.**

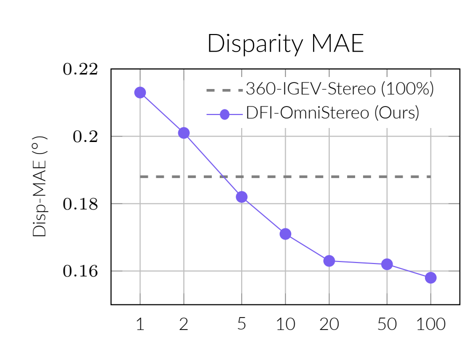

Figure 4. **Training sample-efficient learning analysis.**

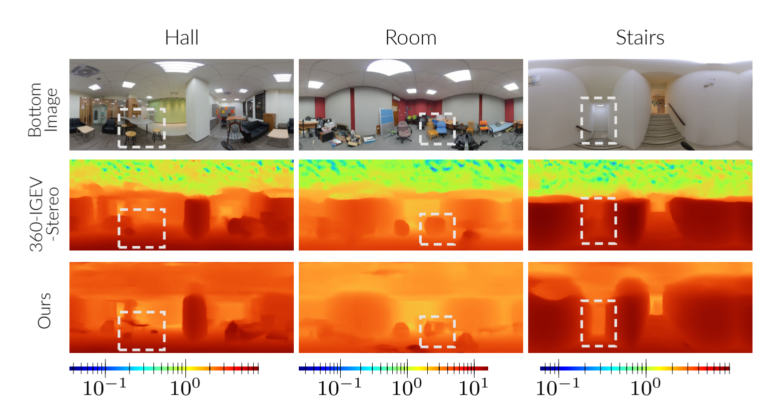

Figure 5. **Qualitative comparison of generalization to real-world images from [5].**

Contributions

- **Depth foundation model** as a **feature extractor** in an **iterative** optimization-based **stereo matching** architecture
- Introduction of a **two-stage training strategy** to adapt the monocular foundation model features to omnidirectional stereo matching
- Scale-invariant error in log space (**SILog loss**) for **stereo matching**
- **SotA results** on the Helvipad dataset [3]
- Promising **generalization capabilities** and **high** training **sample efficiency**

References & Acknowledgments

[1] L. Yang, B. Kang, Z. Huang *et al.*, "Depth Anything V2," in *NeurIPS*, 2024.
[2] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *CVPR*, 2023.
[3] M. Zayene, J. Endres, A. Havolli *et al.*, "Helvipad: A real-world dataset for omnidirectional stereo depth estimation," in *CVPR*, 2025.
[4] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *CVPR*, 2018.
[5] N.-H. Wang, B. Solarte, Y.-H. Tsai *et al.*, "360SD-Net: 360° stereo depth estimation with learnable cost volume," in *ICRA*, 2020.

This project was partially supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008). Additionally, this work has also been co-funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center [LOEWE/1/12/519/03/05.001(0016)/72] and was supported by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223).