PDF-Omni: Poincaré Dual Disk Distortion Field-based Recurrent Update for Omnidirectional Stereo Matching

Yunseok Yang, Eunjin Son, and Sang Jun Lee*

Jeonbuk National University, Jeonju 54896, Korea
{yunseok,eunjinson,sj.lee}@jbnu.ac.kr

Abstract. Omnidirectional stereo matching (OSM) estimates 360° depth from multi-view fisheye images, where inherent fisheye distortion causes sparse matching cues in seam regions. However, existing methods rely on spatially uniform update policies, degrading performance in these challenging regions. To address this limitation, we propose PDF-Omni, a geometry-aware OSM framework built on a Poincaré Dual Disk distortion field that models spatially varying geometric uncertainty across seam regions. We introduce a distortion-informed GRU with distortion-aware attention to selectively expand the receptive field in high-uncertainty regions while preserving efficiency in low-uncertainty ones. We further present a local entropy cumulative loss to supervise high-entropy cost distributions in ambiguous regions. Experimental results demonstrate state-of-the-art performance with consistent gains especially in seam regions. The code is available at https://github.com/ynskyang/PDF-Omni

Keywords: Omnidirectional Stereo Matching · Hyperbolic Geometry · Fisheye Distortion

1 Introduction

Omnidirectional depth estimation provides complete 360° scene understanding, essential for 3D perception in autonomous navigation and driving [24, 29]. Omnidirectional stereo matching (OSM) extends multi-view stereo (MVS) to fisheye images captured by a multi-camera rig, where four cameras face the front, right, back, and left directions. Unlike monocular methods limited by inherent scale ambiguity, OSM leverages explicit geometric triangulation from multiple calibrated viewpoints, enabling metric depth recovery across the surrounding scene.

OSM adopts spherical sweeping [9], warping multi-view fisheye images onto concentric spheres at varying depth hypotheses and constructing a spherical cost volume. Existing works follow two paradigms in volume construction and regularization. The first concatenates multi-view spherical feature volumes into a unified 4D cost volume and regularizes it through a 3D CNN encoder-decoder [25, 27].

* Corresponding author.

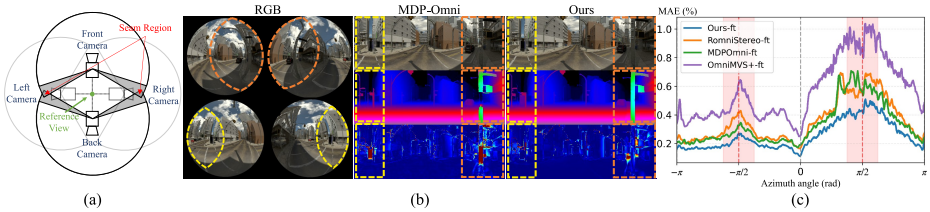

Fig. 1: (a) Four-camera rig where the gray shaded areas between the front and back cameras indicate peripheral and view-transition regions that define seam regions. (b) Predicted depth maps on Sunny: PDF-Omni reduces artifacts in dashed regions compared to MDP-Omni [17]. (c) Azimuth-wise MAE (%) on Sunny for fine-tuned models. Each value is the average over the elevation levels across the test set. Red shaded bands denote seam regions at $\theta = \pm\pi/2$. Error peaks at seam regions are reduced by PDF-Omni.

A recent method further improves the 3D CNN paradigm by introducing cascade coarse-to-fine volume regularization [17]. The second adaptively weights opposite-facing camera pairs to form reference and target volumes, and regularizes them through a RAFT-based iterative update framework [10].

In the adaptive-weighting method [10], the front-back camera pair forms the reference volume, and the context feature and correlation volume are subsequently defined in the reference perspective. The seam regions correspond to the overlap areas of the front and back cameras field of view (FOV), where fisheye radial distortion is most severe. In seam regions, neither camera in the front-back pair provides a geometrically dominant viewpoint, and both cameras observe the scene through their most distorted peripheral fields. The combination of severe radial distortion and ambiguous view assignment makes depth estimation in seam regions substantially more difficult compared to each camera's central region. As shown in Fig. 1, depth errors consistently concentrate at seam regions, corresponding to azimuths of $\theta = \pm\pi/2$ across all existing methods, including both 3D CNN-based and RAFT-based frameworks.

Several recent works have been proposed to address fisheye distortion and matching ambiguity in OSM. Chen et al. [5] incorporate tangent plane sampling and spherical positional encoding to reduce distortion artifacts in feature extraction and cost aggregation. Son et al. [17] propose azimuth-based multi-view volume fusion to mitigate false matches from occlusions. However, these methods overlook the spatially varying uncertainty inherent in seam regions and apply uniform spatial context across all regions. As a result, they lack the capacity to selectively capture broader context where geometric ambiguity is highest. Their supervision also remains insensitive to the high-entropy cost distributions that arise at seam boundaries.

In this paper, we propose PDF-Omni, a geometry-aware framework for omnidirectional stereo matching. PDF-Omni incorporates a Poincaré Dual Disk (PDD) distortion field that assigns a continuous uncertainty measure to every pixel on the equirectangular domain. The PDD distortion field leverages the

boundary-divergent property of Poincaré geodesic distance to model the nonlinear increase of fisheye radial distortion toward seam regions. The PDD distortion field is integrated into a Distortion-Aware Attention (DAA) mechanism that drives a Distortion-informed GRU (DiGRU), an update block that selectively allocates compact kernels to low-uncertainty regions and expanded kernels to high-uncertainty seam regions. Moreover, we introduce a Local Entropy Cumulative (LEC) loss that adapts the supervision range to per-pixel cost volume entropy, enabling the network to learn from high-entropy distributions in ambiguous regions. PDF-Omni demonstrates state-of-the-art performance on five datasets, with consistent gains in seam regions.

In summary, the main contributions are as follows:

- We propose PDF-Omni, a geometry-aware OSM framework that models spatially varying fisheye distortion through a PDD distortion field. The PDD distortion field captures both radial distortion and view-transition ambiguity in a unified formulation.
- We present distortion-informed iterative refinement, comprising DAA and DiGRU, that translates the PDD distortion field into adaptive receptive field allocation for high-uncertainty seam regions.
- We introduce a LEC loss that adapts the supervision range to per-pixel cost volume entropy, reducing gradient noise from ambiguous correspondences.

2 Related Work

2.1 Omnidirectional Stereo Matching

Early OSM pipelines relied on spherical sweeping and SGM-based aggregation for 360° depth estimation [8, 9, 26]. SweepNet [26] introduced CNN-based cost computation on spherical image pairs but retained SGM for cost aggregation. OmniMVS [25] enabled end-to-end learning by replacing SGM with 3D CNN-based cost regularization. OmniMVS+ [27] reduced memory consumption through an interleaved sweeping strategy. Subsequent works improved OSM along several directions. Efficiency-oriented methods achieved real-time inference through fast cost aggregation, lightweight architectures, and cascade sweeping [14,20,22]. Un-OmniMVS [4] reduced reliance on ground-truth depth by using pseudo-stereo supervision from back-to-back fisheye pairs. To improve robustness to distortion and occlusion, prior work employed tangent plane sampling with spherical positional encoding [5] and azimuth-dependent fusion weights [17]. RomniStereo [10] introduced iterative refinement for OSM by forming paired reference and target volumes and updating them with an opposite adaptive weighting scheme. Despite these advances, current OSM methods do not explicitly model spatially varying uncertainty at seam regions, leaving selective receptive field allocation unexplored.

2.2 Iterative Refinement for Stereo Matching

Prior to iterative approaches, one-shot methods constructed cost volumes and applied global regularization in a single pass [3, 7, 11, 31]. The one-shot formulation achieves strong accuracy but incurs substantial memory and computational cost. RAFT [19] addressed this limitation in optical flow by replacing one-shot estimation with iterative GRU-based refinement over a 4D all-pairs correlation volume. RAFT-Stereo [13] extended this paradigm to binocular stereo matching, and subsequent works refined it further through cascade recurrent updates [12] and hybrid cost volumes [30]. RAFT-based methods apply a uniform convolutional GRU update to every pixel. Selective-Stereo [21] addressed the resulting limitation by blending small and large-kernel GRU updates through frequency-based spatial attention. In OSM, spatially varying difficulty arises primarily from geometric uncertainty due to fisheye distortion and view transitions, rather than frequency content. Inspired by Selective-Stereo [21], we extend the selective refinement strategy to OSM by replacing frequency-based attention with a Poincaré distortion field, enabling geometry-driven receptive field allocation.

2.3 Hyperbolic Geometry for Fisheye Modeling

Hyperbolic representation learning computes embeddings in non-Euclidean spaces with constant negative curvature, increasing representation capacity for hierarchical data. Nickel and Kiela [16] introduced Poincaré embeddings that optimize representations in the Poincaré ball. The exponential volume growth toward the boundary enables compact encodings that capture hierarchy and similarity more efficiently compared to Euclidean embeddings. FisheyeHDK [1] connected boundary expansion in the Poincaré ball to fisheye optics and learned distortion-aware deformable kernel positions through hyperbolic feature transformations without explicit calibration. DCPB [23] formalized the correspondence between top-view fisheye projection and the Poincaré disk, and derived kernel offsets from geodesic distances to improve segmentation on distorted imagery. However, both methods focus on single-camera recognition and do not extend to multi-camera OSM, in which distortion varies jointly with azimuth and view-transition geometry. We address this limitation by instantiating two Poincaré disks at the front and back camera centers and modeling seam region uncertainty in a unified field.

3 Method

PDF-Omni follows an adaptive iterative refinement strategy consisting of three stages: feature extraction, volume generation, and iterative refinement, as illustrated in Fig. 2. The subsequent sections are organized as follows: Section 3.1 describes the overall pipeline, Section 3.2 presents the Poincaré Dual Disk distortion field, Section 3.3 describes distortion-informed iterative refinement, and Section 3.4 details the loss function.

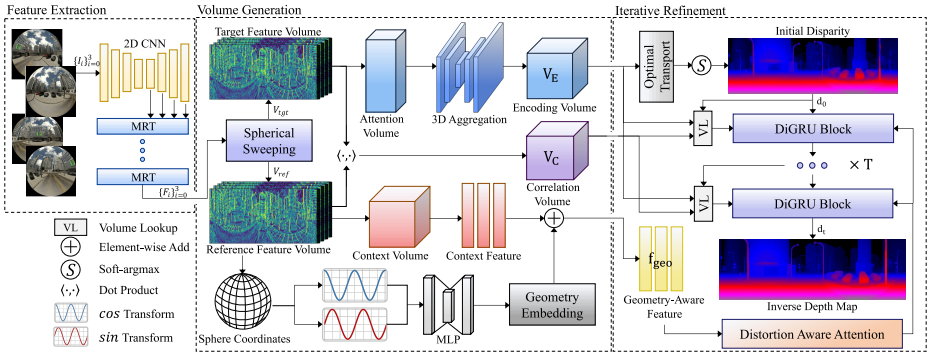

Fig. 2: Overview of PDF-Omni. A shared 2D CNN with multi-resolution transformers extracts multi-view features. Spherical sweeping constructs reference and target volumes, from which an encoding volume $\mathbf{V}_E$ and a correlation volume $\mathbf{V}_C$ are derived. Optimal transport with softargmax regression produces the initial depth $\mathbf{d}_0$. DAA converts the PDD distortion field into a spatial attention map, and DiGRU iteratively refines the depth prediction through distortion-modulated recurrent updates.

3.1 Overall Pipeline

Feature Extraction. A shared 2D CNN processes four fisheye images $\{I_i\}_{i=0}^{3}$ from front, right, back, and left cameras, extracting multi-scale feature maps at strides of 2, 4, 8, and 16. Multi-Resolution Transformers [15] refine the multi-scale features through cross-attention and self-attention across scales, yielding multi-view feature maps $\{F_i\}_{i=0}^{3} \in \mathbb{R}^{C\times\frac{H_I}{2}\times\frac{W_I}{2}}$.

Volume Generation. The volume generation stage follows spherical sweeping [9]. Each feature map F_i is projected onto N concentric spheres indexed by inverse depth and warped on the equirectangular planes, yielding a spherical feature volume for each view. Following Jiang et al. [10], adaptive weighting combines opposite-facing volumes into a reference volume V_{ref} and a target volume V_{tgt}. From the paired volumes, an attention volume is first constructed through groupwise correlation between V_{ref} and V_{tgt}. A lightweight 3D hourglass network aggregates the attention volume into an encoding volume $\mathbf{V}_E$ that captures non-local geometric context. An initial inverse depth estimate $\mathbf{d}_0$ is then obtained from $\mathbf{V}_E$ by applying entropy-regularized optimal transport [6] followed by softargmax regression. In parallel, a correlation volume $\mathbf{V}_C$ is computed as a dot-product correlation between V_{ref} and V_{tgt} across the inverse depth dimension. Both $\mathbf{V}_E$ and $\mathbf{V}_C$ are organized into multi-level pyramids by progressively halving the inverse depth dimension, providing coarse-to-fine evidence for iterative refinement. The reference volume additionally serves as a context volume. A context branch aggregates features along the inverse depth axis to produce a 2D context feature. Fourier-mapped spherical coordinates [18] are processed by a multilayer perceptron to obtain a geometry embedding, which is added to the context feature to form the geometry-aware feature $\mathbf{f}_{\text{geo}}$.

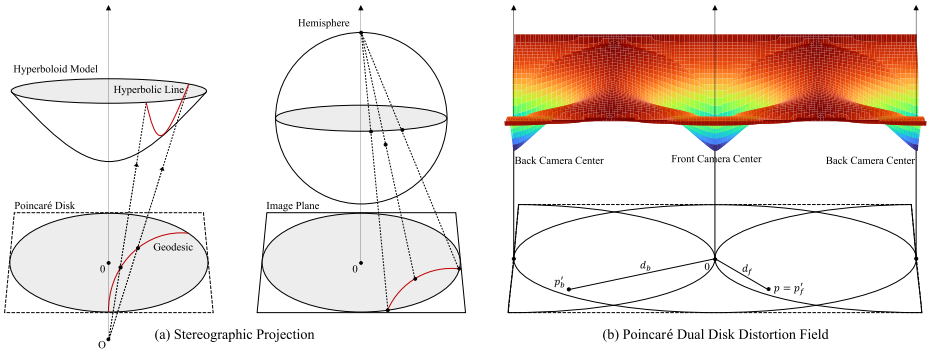

Fig. 3: (a) Stereographic projection from the hyperboloid model to the Poincaré disk (left). Analogous stereographic projection from the hemisphere to the fisheye image plane (right). (b) Poincaré dual disk distortion field $\mathcal{D}$ on the equirectangular domain. The back camera optical axis appears at both horizontal boundaries due to azimuth periodicity.

Iterative Refinement. The iterative refinement stage starts from $\mathbf{d}_0$ and performs T update iterations. At each iteration t, volume lookup retrieves local features from each pyramid level of $\mathbf{V}_E$ and $\mathbf{V}_C$ around the current estimate $\mathbf{d}_{t-1}$. The retrieved multi-level features are concatenated with $\mathbf{d}_{t-1}$ and projected through a convolutional layer into a compact update representation. The geometry-aware feature $\mathbf{f}_{\text{geo}}$ is concatenated with this representation. DAA converts the PDD distortion field $\mathcal{D}$ into a spatial attention map $\mathbf{A}$, which guides the DiGRU block to predict an inverse depth residual $\Delta\mathbf{d}_t$ through distortion-modulated recurrent updates. The estimate is updated as $\mathbf{d}_t = \mathbf{d}_{t-1} + \Delta\mathbf{d}_t$. After T iterations, a learned convex upsampling mask [13, 19] upsamples the final inverse depth map to the input resolution.

3.2 Poincaré Dual Disk Distortion Field

Fisheye lenses produce radially increasing distortion, with mild warping near the optical axis and strong compression at the periphery. Fig. 3(a) illustrates the structural analogy between the stereographic projection from the hyperboloid model onto the Poincaré disk and the projection from the hemisphere onto the fisheye image plane. Both projections share a boundary-divergent property in that the metric distance grows nonlinearly toward the boundary. Prior works [1, 23] exploited this property to model fisheye radial distortion through Poincaré geodesic distance. We extend the single-disk formulation to a dual-disk construction that covers the full 360° omnidirectional domain.

Poincaré Disk Model. The Poincaré ball model [2] is defined by the Riemannian manifold $(\mathbb{D}_c^d, g^{\mathbb{D}_c})$, where $\mathbb{D}_c^d = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\| < 1/\sqrt{c}\}$ is an open ball of radius $1/\sqrt{c}$ with constant negative curvature $-c$. In this work, we set $c = 1$ and $d = 2$, yielding the unit Poincaré disk $\mathbb{D}^2 = \{\mathbf{x} \in \mathbb{R}^2 \mid \|\mathbf{x}\| < 1\}$ that maps

directly onto the normalized equirectangular coordinate domain. The Poincaré metric tensor $g^{\mathbb{D}}_{\mathbf{x}}$ is defined as

$$g^{\mathbb{D}}_{\mathbf{x}} = \lambda^2_{\mathbf{x}}\, g^E, \qquad \lambda_{\mathbf{x}} := \frac{2}{1 - \|\mathbf{x}\|^2}, \tag{1}$$

where g^E denotes the Euclidean metric tensor. The conformal factor $\lambda_{\mathbf{x}}$ rescales the Euclidean metric to account for the curvature of the hyperbolic space. Building upon this metric, the geodesic distance $d_{\mathbb{D}}$ from the origin to $\mathbf{p}' \in \mathbb{D}^2$ is defined as

$$d_{\mathbb{D}}\big(\mathbf{0},\, \mathbf{p}'\big) = \cosh^{-1}\!\left(1 + \frac{2\,\|\mathbf{p}'\|^2}{1 - \|\mathbf{p}'\|^2}\right). \tag{2}$$

The distance $d_{\mathbb{D}}$ diverges toward the disk boundary and provides a monotonic measure of radial distortion magnitude with respect to $\|\mathbf{p}'\|$. To apply the geodesic distance to the equirectangular domain, normalized coordinates $(x, y) \in [-1, 1]^2$ are defined, where x corresponds to azimuth and y corresponds to elevation. Each coordinate $\mathbf{p} = (x, y)$ is clamped to the disk interior to ensure numerical stability:

$$\mathbf{p}' = \begin{cases} (1 - \epsilon)\,\dfrac{\mathbf{p}}{\|\mathbf{p}\|}, & \|\mathbf{p}\| \geq 1, \\ \mathbf{p}, & \text{otherwise}, \end{cases} \tag{3}$$

where $\epsilon > 0$ prevents instability near the disk boundary.

Dual Disk Construction. A Poincaré disk is defined for each of the front and back cameras by centering the disk at the respective optical axis. The front and back camera axes point in opposite directions on the unit sphere of viewing directions, yielding two complementary distortion profiles on the equirectangular domain. For the front and back cameras, the geodesic distances are

$$d_f = d_{\mathbb{D}}\big(\mathbf{0},\, \mathbf{p}'_f\big), \quad d_b = d_{\mathbb{D}}\big(\mathbf{0},\, \mathbf{p}'_b\big), \tag{4}$$

where $\mathbf{p}'_f$ denotes the clamped Poincaré disk coordinate derived from (x, y) using Eq. (3). The back disk applies an azimuth shift of half a period relative to the front disk. The shifted coordinate is $\mathbf{p}_b = \big(((x + 2) \bmod 2) - 1,\ y\big)$, and $\mathbf{p}'_b$ is obtained by applying Eq. (3) to $\mathbf{p}_b$. The dual disk construction covers the full 360° azimuth range while preserving a monotonically increasing mapping from radial position to distortion magnitude for each of the two reference cameras.

Poincaré Dual Disk Distortion Field. The dual disk distances d_f and d_b are combined into a single scalar field $\mathcal{D}\colon [-1, 1]^2 \to [0, 1)$ that quantifies the geometric uncertainty at each pixel. The field consists of two terms: a radial term that measures baseline distortion magnitude, and a view transition term that identifies seam regions.

The radial term selects the geodesic distance from the closer camera center at each pixel and normalizes it to $[0, 1)$:

$$\bar{d}(x, y) = \tanh\big(\min(d_f,\, d_b)\big). \tag{5}$$

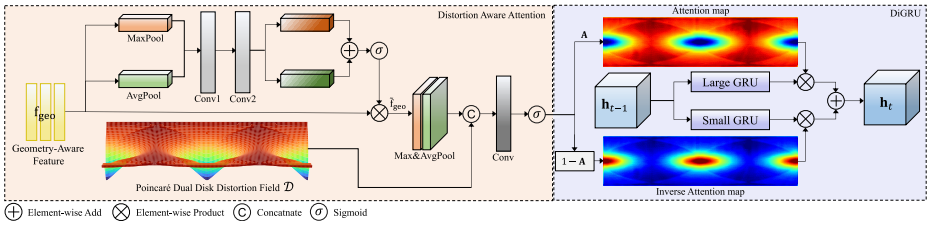

Fig. 4: Architecture of DAA and DiGRU. DAA produces a spatial attention map **A** from the PDD distortion field $\mathcal{D}$. DiGRU blends a small-kernel GRU branch with a large-kernel GRU branch using **A**, selectively enlarging the receptive field in high-distortion seam regions.

Pixels near a camera center yield small $\bar{d}$ values, indicating mild distortion, while peripheral pixels yield large values.

The view transition term measures whether a pixel lies near the boundary between the two cameras:

$$w_{vt}(x,y) = 1 - \frac{|d_f - d_b|}{d_f + d_b + \epsilon}\,. \tag{6}$$

When d_f and d_b are similar, neither camera provides a dominant viewpoint and w_{vt} approaches 1, indicating a seam region. When one camera is dominant, w_{vt} approaches 0.

The final field combines both terms so that the distortion value is amplified only for pixels that are both peripheral and located at a camera boundary:

$$\mathcal{D}(x,y) = \tanh\big(\bar{d}(x,y) + \alpha \cdot w_{vt}(x,y) \cdot \bar{d}(x,y)\big), \tag{7}$$

where $\alpha > 0$ controls the emphasis on view transition regions. The additive term $\alpha \cdot w_{vt} \cdot \bar{d}$ is significant only in seam regions where both $\bar{d}$ and w_{vt} are large, leaving non-seam regions unaffected. The field $\mathcal{D}$ is precomputed once per spatial resolution and reused across refinement iterations.

3.3 Distortion-Informed Iterative Refinement

In seam regions, front-back cameras observe the scene through their most distorted peripheral fields, resulting in sparse local matching cues that demand broader spatial context for reliable updates. In contrast, non-seam regions, where distortion is mild, benefit from compact local updates. To address this disparity, the refinement stage maps the PDD distortion field $\mathcal{D}$ to pixel-wise receptive field control, as illustrated in Fig. 4. Specifically, DAA predicts a spatial attention map **A** from the PDD distortion field. DiGRU blends small-kernel and large-kernel updates conditioned on **A**.

Distortion-Aware Attention. DAA derives from the CBAM attention mechanism [28] and incorporates the PDD distortion field $\mathcal{D}$ as an explicit geometric input. Channel attention recalibrates $\mathbf{f}_{\text{geo}}$ into $\tilde{\mathbf{f}}_{\text{geo}}$ by reweighting channels

based on global average-pooled and max-pooled statistics. In the spatial attention stage, channel-wise average pooling and max pooling over $\tilde{\mathbf{f}}_{\text{geo}}$ are concatenated with $\mathcal{D}$, and a $k\times k$ convolution followed by a sigmoid activation produces the distortion-aware attention map

$$\mathbf{A} = \sigma\big(\text{Conv}_{k\times k}([\text{AvgPool}_c(\tilde{\mathbf{f}}_{\text{geo}}),\ \text{MaxPool}_c(\tilde{\mathbf{f}}_{\text{geo}}),\ \mathcal{D}])\big). \tag{8}$$

By conditioning on $\mathcal{D}$, DAA assigns high attention to seam-region pixels based on geometric uncertainty rather than relying solely on appearance cues.

Distortion-informed GRU. To adapt the receptive field to local geometric uncertainty, DiGRU maintains two GRU branches with different kernel sizes and blends their outputs using the attention map $\mathbf{A}$. The small-kernel branch uses 1×1 convolutions for efficient per-pixel updates in low-distortion regions, while the large-kernel branch employs separable $1\times K$ and $K\times1$ convolutions to aggregate wider spatial context for seam regions. The update rule is

$$\mathbf{h}_t = (1-\mathbf{A})\odot \text{GRU}_{\text{s}}(\mathbf{h}_{t-1},\, \mathbf{x}_t)\ +\ \mathbf{A}\odot \text{GRU}_{\text{l}}(\mathbf{h}_{t-1},\, \mathbf{x}_t), \tag{9}$$

where $\mathbf{x}_t = [\tilde{\mathbf{f}}_{\text{geo}},\ \mathbf{m}_t]$ concatenates the recalibrated geometry-aware feature and a motion feature $\mathbf{m}_t$. The motion feature is encoded from local cost volume evidence and the current estimate $\mathbf{d}_{t-1}$. A convolutional head predicts an inverse depth residual $\Delta\mathbf{d}_t$ from $\mathbf{h}_t$, and the estimate is updated as $\mathbf{d}_t = \mathbf{d}_{t-1} + \Delta\mathbf{d}_t$.

3.4 Loss Function

Local Entropy Cumulative Loss. Previous methods supervise only the regressed depth value, leaving the shape of the depth probability distribution unconstrained. The LEC loss directly regularizes the probability volume $\mathbf{P} \in \mathbb{R}^{N\times H\times W}$ by adapting a local supervision radius to the prediction uncertainty at each pixel. Pixels with higher entropy are assigned a larger neighborhood radius, which relaxes supervision for ambiguous predictions. For each valid pixel (i,j), the entropy of the depth distribution $\{p_n^{(i,j)}\}_{n=0}^{N-1}$ is normalized by the maximum possible entropy and clamped to $[0,1]$ for numerical stability:

$$\tilde{H}(i,j) = \text{clip}\left(\frac{-\sum_{n=0}^{N-1} p_n^{(i,j)} \log\big(p_n^{(i,j)}+\epsilon\big)}{\log N},\, 0,\, 1\right), \tag{10}$$

where $\text{clip}(x,l,u) = \min(\max(x,l),u)$. The neighborhood radius $r(i,j)$ is determined by the normalized entropy as

$$r(i,j) = \max\big(1,\ \lceil \tilde{H}(i,j)\cdot(N-1)\rceil\big), \tag{11}$$

where $\lceil\cdot\rceil$ denotes the ceiling function. Let $n_{\text{gt}}^{(i,j)} \in \{0,\dots,N-1\}$ denote the ground-truth depth index at pixel (i,j). The neighborhood around $n_{\text{gt}}^{(i,j)}$ is defined as

$$\mathcal{N}(i,j) = \big\{n \in \{0,\dots,N-1\} \mid |n - n_{\text{gt}}^{(i,j)}| \le r(i,j)\big\}. \tag{12}$$

The local cumulative probability within the adaptive neighborhood is

$$m(i,j) = \sum_{n \in \mathcal{N}(i,j)} p_n^{(i,j)}. \tag{13}$$

The LEC loss penalizes insufficient local cumulative probability through a hinge formulation

$$\mathcal{L}_{\text{lec}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \big[1 - m(i,j)\big]_+, \tag{14}$$

where $[\cdot]_+ = \max(0, \cdot)$ and $\mathcal{M}$ is the valid pixel set. The radius $r(i,j)$ is computed from $\tilde{H}(i,j)$ with a stop-gradient operation, preventing the network from inflating entropy to artificially widen the supervision tolerance.

Total Loss. The sequence loss supervises all T intermediate depth estimates with exponentially increasing weights [10, 19]:

$$\mathcal{L}_{\text{seq}} = \sum_{t=1}^{T} \gamma^{T-t} \left\| \mathbf{d}_{\text{gt}} - \mathbf{d}_t \right\|_1. \tag{15}$$

Finally, we define the total loss as follows:

$$\mathcal{L} = \mathcal{L}_{\text{seq}} + \lambda_{\text{lec}}\, \mathcal{L}_{\text{lec}}. \tag{16}$$

4 Experiments

4.1 Implementation Details

We implement PDF-Omni in PyTorch and conduct all experiments on a single NVIDIA RTX 6000 Ada GPU. The base feature channel dimension is set to $C = 32$ for fair comparison with previous methods [10, 17, 27] which adopt the same channel size. The number of spherical sweeping hypotheses is $N = 192$. For the geometry embedding, the Fourier feature scale is $\sigma = 1$. In the PDD distortion field, the view transition weighting factor is $\alpha = 0.5$ (Eq. (7)). DiGRU uses a 1×1 convolution branch and a separable 1×7 + 7×1 convolution branch, with $T = 12$ refinement iterations during both training and evaluation, consistent with RomniStereo [10]. The sequence loss discount factor is $\gamma = 0.9$ and LEC loss weight is $\lambda_{\text{lec}} = 0.1$ (Eq. (16)). Following prior works [10, 17, 25, 27], we pre-train PDF-Omni on OmniThings for 30 epochs using the AdamW optimizer with a one-cycle learning rate schedule and a maximum learning rate of 5×10^{-4}, and fine-tune for 15 epochs on OmniHouse and Sunny.

4.2 Datasets and Evaluation Metrics

Datasets. We evaluate on five synthetic datasets introduced by Won *et al.* [25–27]. Each scene consists of four fisheye images captured by cameras with a 220° FOV oriented to the front, right, back, and left, with an input resolution of

Table 1: Quantitative comparison on OmniThings, OmniHouse, and Sunny. Best results are in **bold** and second best are underlined. Lower is better for all metrics.

Method	OmniThings					OmniHouse					Sunny				
	>1	>3	>5	MAE	RMS	>1	>3	>5	MAE	RMS	>1	>3	>5	MAE	RMS
Non-learning based method															
Sphere-Stereo [14]	80.01	56.67	44.06	9.14	14.06	65.84	27.29	12.84	2.82	4.60	76.46	45.99	28.46	4.92	8.35
Trained on OmniThings only															
OmniMVS [25]	47.72	15.12	8.91	2.40	5.27	30.53	10.29	6.27	1.72	4.05	27.16	6.13	3.98	1.24	3.09
S-OmniMVS [5]	28.03	10.40	6.33	1.48	3.68	18.86	8.05	4.90	1.06	2.41	17.19	6.03	3.89	1.11	3.60
OmniMVS$^{+}_{32}$ [27]	20.70	8.18	5.49	1.37	4.11	19.89	5.89	3.99	1.30	2.64	13.57	4.81	3.10	0.88	2.56
RomniStereo$_{32}$ [10]	20.42	8.49	5.81	1.39	4.22	12.13	4.73	**3.02**	**0.80**	**1.85**	12.28	5.59	3.79	0.80	2.68
MDP-Omni [17]	18.37	7.09	4.59	1.20	3.79	17.13	7.20	4.68	1.16	2.62	12.72	4.97	3.26	0.79	2.62
Ours	**14.86**	**6.34**	**4.21**	**1.08**	**3.66**	**8.07**	**3.89**	3.03	0.86	2.23	**8.61**	**3.88**	**2.73**	**0.61**	**2.42**
Fine-tuned on OmniHouse and Sunny															
OmniMVS-ft [25]	50.28	22.78	15.60	3.52	7.44	21.09	4.63	2.58	1.04	1.97	13.93	2.87	1.71	0.79	2.12
S-OmniMVS-ft [5]	-	-	-	-	-	6.99	1.79	0.97	0.42	1.06	6.66	2.18	1.40	0.47	1.98
OmniMVS$^{+}_{32}$-ft [27]	44.79	27.17	20.41	4.23	8.42	9.70	3.51	2.13	0.64	1.69	7.48	3.57	2.42	0.57	2.42
RomniStereo$_{32}$-ft [10]	34.32	19.76	14.22	2.81	6.47	6.02	2.49	1.73	0.49	1.31	5.19	1.98	1.23	0.36	1.55
MDP-Omni-ft [17]	29.53	14.05	9.12	1.96	5.08	5.61	2.19	1.50	0.44	1.15	4.51	1.43	0.86	0.33	1.43
Ours-ft	**17.75**	**8.12**	**5.51**	**1.31**	**4.17**	**2.86**	**0.96**	**0.58**	**0.23**	**0.79**	**3.54**	**1.32**	**0.80**	**0.25**	**1.26**

$H_I = 768$, $W_I = 800$. OmniThings provides 9,216 training and 1,024 test scenes of generic objects placed in diverse photometric and geometric environments. OmniHouse provides 2,048 training and 512 test scenes of realistic indoor environments rendered from 451 house models. Sunny, Cloudy, and Sunset each contain 700 training and 300 test scenes of outdoor driving under the different weather conditions.

Evaluation metrics. Following OmniMVS+ [27], the ground-truth inverse depth maps are cropped to an elevation range of $-\pi/4 \leq \phi \leq \pi/4$ and an azimuth range of $-\pi \leq \theta \leq \pi$, yielding an evaluation resolution of $H = 160$, $W = 640$. We report the percentage of pixels whose error exceeds thresholds 1, 3, and 5 (denoted >1, >3, >5), mean absolute error (MAE), and root mean squared error (RMS). All errors are computed as the absolute difference between predicted and ground-truth inverse depth indices along the depth hypothesis dimension, normalized by the total number of hypotheses N and expressed as a percentage. Lower values indicate better performance for all metrics. For seam-region evaluation, we define two seam bands of width $\pi/4$ centered at $\theta = \pm\pi/2$, corresponding to the front-back camera boundaries. Seam metrics are computed as the elevation-axis average across the test set within the seam bands, and non-seam metrics are computed over the remaining azimuth range in the same manner.

4.3 Experimental Results

Table 1 and Table 2 present quantitative comparisons under two training protocols: pre-training on OmniThings and fine-tuning on OmniHouse and Sunny.

Table 2: Quantitative comparison on Cloudy and Sunset.

Method	Cloudy					Sunset				
	>1	>3	>5	MAE	RMS	>1	>3	>5	MAE	RMS
Non-learning based method										
Sphere-Stereo [14]	77.57	47.08	28.39	4.50	7.21	77.38	46.11	28.49	5.15	8.89
Trained on OmniThings only										
OmniMVS [25]	28.13	5.37	3.54	1.17	2.83	26.70	6.19	4.02	1.24	3.06
OmniMVS$^{+}_{32}$ [27]	13.59	4.81	3.15	0.87	2.53	13.36	4.71	2.93	0.87	2.50
RomniStereo$_{32}$ [10]	11.86	5.08	3.44	0.75	2.50	12.30	5.45	3.48	0.78	2.67
Ours	**8.91**	**3.93**	**2.78**	**0.61**	**2.33**	**8.51**	**3.68**	**2.47**	**0.56**	**2.18**
Fine-tuned on OmniHouse and Sunny										
OmniMVS-ft [25]	12.20	2.48	1.46	0.72	1.85	14.14	2.88	1.71	0.79	2.04
OmniMVS$^{+}_{32}$-ft [27]	7.29	3.38	2.30	0.54	2.31	7.82	3.60	2.42	0.58	2.36
RomniStereo$_{32}$-ft [10]	5.63	2.03	1.29	0.39	1.72	5.53	2.13	1.34	0.37	1.61
MDP-Omni-ft [17]	5.34	1.68	1.03	0.36	1.53	4.90	1.51	0.90	0.35	1.46
Ours-ft	**3.92**	**1.37**	**0.84**	**0.27**	**1.32**	**3.93**	**1.43**	**0.85**	**0.27**	**1.30**

Table 3: Seam and non-seam region quantitative comparison on Sunny.

Method	Region	Sunny				
		>1	>3	>5	MAE	RMS
OmniMVS$^{+}_{32}$-ft [27]	Seam	8.955	4.191	2.957	0.719	3.214
	Non-Seam	6.977	3.354	2.232	0.524	2.244
RomniStereo$_{32}$-ft [10]	Seam	6.618	2.757	1.829	0.475	2.369
	Non-Seam	4.714	1.726	1.033	0.318	1.628
MDP-Omni-ft [17]	Seam	5.879	1.915	1.206	0.447	2.706
	Non-Seam	4.050	1.267	0.745	0.294	1.519
Ours-ft	Seam	**4.314**	**1.730**	**1.119**	**0.332**	**1.984**
	Non-Seam	**3.270**	**1.173**	**0.687**	**0.229**	**1.258**

Pre-trained Models. When trained on OmniThings only, PDF-Omni achieves the best performance across all five metrics on OmniThings. On the cross-domain datasets Sunny, Cloudy, and Sunset, PDF-Omni also achieves the lowest errors on all metrics without any exposure to outdoor driving data during training. On OmniHouse, PDF-Omni achieves the lowest >1 and >3 errors, demonstrating strong generalization to indoor environments.

Fine-tuned Models. After fine-tuning, PDF-Omni-ft achieves the best results across all five datasets on every metric. The improvement is particularly notable on OmniThings, where PDF-Omni-ft reduces >1 error from 29.53 to 17.75 compared to MDP-Omni-ft [17], demonstrating that the proposed distortion-aware refinement and entropy-adaptive supervision jointly strengthen depth estimation. Consistent gains on Cloudy and Sunset further show robustness to varying weather conditions.

Seam Region Analysis. Fine-tuned models are evaluated separately on seam and non-seam regions of Sunny in Table 3. PDF-Omni-ft achieves the lowest error in both regions, and the performance gap between seam and non-seam is the smallest across all metrics among all compared methods. Specifically, the >1 gap of PDF-Omni-ft is 1.04 percentage points, compared to 1.83 for MDP-Omni-ft and 1.90 for RomniStereo$_{32}$-ft. The reduced disparity shows that distortion-informed receptive field modulation effectively mitigates the geometric ambiguity concentrated at seam boundaries without sacrificing non-seam accuracy.

Qualitative Results. Fig. 5 presents qualitative comparisons of fine-tuned models across OmniThings, OmniHouse, Sunny, and real-scene data. PDF-Omni-ft produces sharper depth boundaries and fewer artifacts in seam regions compared to MDP-Omni-ft and RomniStereo$_{32}$-ft, aligning with the quantitative results in Table 1 and Table 3.

Efficiency. Inference memory and runtime are measured on an NVIDIA GeForce RTX 3090. PDF-Omni requires 5,284 MB of GPU memory and 0.36 s per frame. Compared to OmniMVS$^{+}_{32}$ [27], which requires 6,284 MB and 0.62 s, PDF-Omni consumes 15.9% less memory and runs 41.9% faster while delivering substantially higher accuracy.

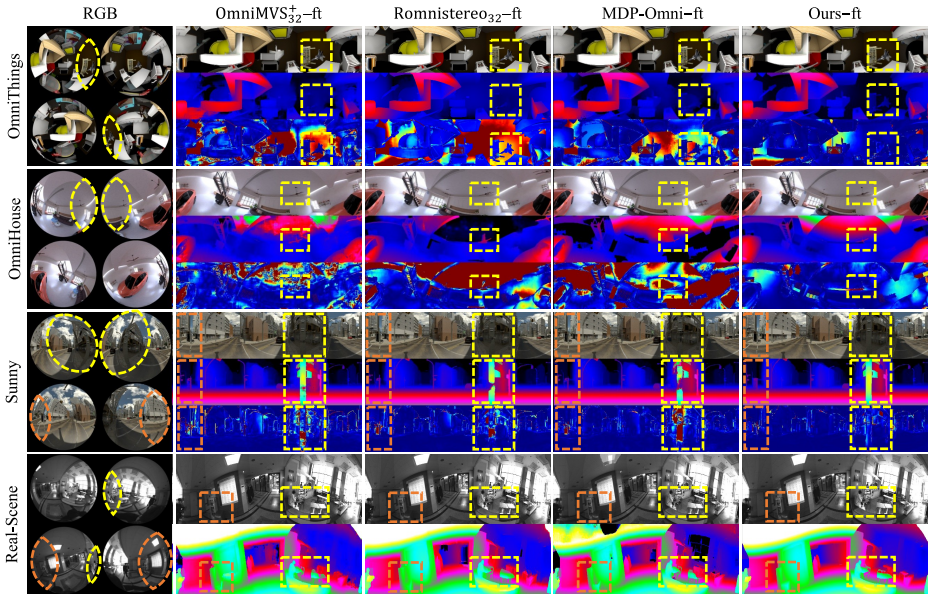

Fig. 5: Qualitative comparison of fine-tuned models.

Table 4: Component ablation on OmniThings, OmniHouse, and Sunny. Each row corresponds to a different combination of DAA+DiGRU, LEC loss, and geometry embedding (GE).

DAA+DiGRU	LEC Loss	GE	OmniThings					OmniHouse					Sunny				
			>1	>3	>5	MAE	RMS	>1	>3	>5	MAE	RMS	>1	>3	>5	MAE	RMS
×	×	×	16.52	7.00	4.67	1.16	3.73	9.93	4.74	3.58	0.92	2.21	9.41	4.27	3.04	0.66	2.54
✓	×	×	16.04	6.85	4.39	1.12	3.69	9.39	4.17	3.22	0.89	**2.10**	**8.19**	**3.64**	**2.51**	**0.61**	2.66
✓	×	✓	15.89	6.80	**4.19**	**1.05**	**3.53**	10.10	4.76	3.58	0.96	2.45	10.23	4.42	2.87	0.77	3.52
✓	✓	×	15.70	6.73	4.71	1.15	3.76	9.50	4.84	3.83	0.99	2.50	9.85	4.56	3.15	0.67	2.56
✓	✓	✓	**14.86**	**6.34**	4.21	1.08	3.66	**8.07**	**3.89**	**3.03**	**0.86**	2.23	8.61	3.88	2.73	**0.61**	**2.42**

4.4 Ablation Studies

We pre-train all ablation models on OmniThings for 30 epochs and evaluate on OmniThings, OmniHouse, and Sunny without fine-tuning.

Component Contributions. The effect of each component is examined in Table 4. The full model combining DAA+DiGRU, LEC loss, and geometry embedding (GE) achieves the best overall balance across all three datasets, with the lowest >1 error on OmniThings and OmniHouse, and the lowest MAE on OmniHouse and Sunny. Individual components provide complementary benefits: DAA+DiGRU reduces >1 error on Sunny from 9.41 to 8.19 through distortion-guided receptive field allocation, while LEC loss improves >3 error on OmniThings from 6.85 to 6.73 by regularizing ambiguous cost distributions. However, combining GE and DAA+DiGRU without LEC loss degrades performance on OmniHouse and Sunny below the baseline, with RMS on Sunny rising from

Table 5: Recurrent unit and DAA variant comparison on OmniThings, OmniHouse, and Sunny. GRU: convolutional GRU; SRU: Selective Recurrent Unit [21]; Eucl.: DAA with Euclidean-based distortion field; Inv.: DAA with inverted distortion map $1 - \mathcal{D}$.

Model	RNN		DAA Variant			OmniThings					OmniHouse					Sunny				
	GRU	SRU	DiGRU	Eucl.	Inv.	Full	>1	>3	>5	MAE	RMS	>1	>3	>5	MAE	RMS	>1	>3	>5	MAE	RMS
Baseline	✓						16.52	7.00	4.67	1.16	3.73	9.93	4.74	3.58	0.92	2.21	9.41	4.27	3.04	0.66	2.54
Selective-Stereo [21]		✓					17.03	7.45	4.69	1.15	3.70	9.30	4.30	3.22	0.82	2.11	10.25	4.26	2.71	0.69	2.85
DiGRU+DAA(Eucl.)			✓	✓			15.75	6.62	**4.19**	**1.07**	**3.58**	8.58	4.14	3.19	**0.81**	**1.98**	14.44	5.92	3.32	0.78	2.78
DiGRU+DAA(Inv.)			✓		✓		15.90	6.95	4.51	1.13	3.73	9.21	4.38	3.23	0.84	2.18	8.89	**3.85**	**2.54**	0.63	2.73
Full Model			✓			✓	**14.86**	**6.34**	4.21	1.08	3.66	**8.07**	**3.89**	**3.03**	0.86	2.23	**8.61**	3.88	2.73	**0.61**	**2.42**

Table 6: Drop-in replacement of the recurrent unit in RomniStereo$_{32}$ [10].

Model	GRU	SRU	DAA+DiGRU	OmniThings					OmniHouse					Sunny				
				>1	>3	>5	MAE	RMS	>1	>3	>5	MAE	RMS	>1	>3	>5	MAE	RMS
RomniStereo$_{32}$ [10]	✓			20.42	8.49	5.81	1.39	4.22	12.13	4.73	3.02	0.80	1.85	12.28	5.59	3.79	0.80	2.68
+Selective-Stereo [21]		✓		18.80	8.23	5.51	1.33	4.15	10.52	4.13	2.71	0.72	1.74	11.82	5.18	3.49	0.75	2.53
+Ours			✓	**17.27**	**7.41**	**4.96**	**1.22**	**4.00**	**9.08**	**3.69**	**2.39**	**0.63**	**1.62**	**11.05**	**4.93**	**3.31**	**0.70**	**2.39**

2.54 to 3.52, as the stronger geometric prior leads to overconfident updates in regions with inherently ambiguous cost volumes. Adding LEC loss resolves this instability by constraining the cost distribution shape, enabling the geometric modules to operate on more reliable probability estimates.

Recurrent Unit and Distortion Field Variants. Recurrent unit designs and DAA distortion field formulations are compared in Table 5. Replacing the baseline ConvGRU with the SRU from Selective-Stereo [21] provides no consistent improvement, with >1 error on OmniThings increasing from 16.52 to 17.03, indicating that the frequency-based modulation strategy designed for pinhole stereo does not transfer effectively to the omnidirectional domain. DiGRU with the full Poincaré-based DAA achieves the best overall balance. The Euclidean variant achieves a lower RMS of 1.98 on OmniHouse but suffers severe degradation on Sunny with >1 error of 14.44, suggesting that the linear distance metric fails to capture nonlinear radial compression in outdoor scenes. The inverted variant shows marginal gains on Sunny but degrades OmniThings and OmniHouse, and the full model achieves the most consistent accuracy across all datasets without such dataset-specific trade-offs.

Generalizability to Existing Architectures. DiGRU+DAA is evaluated as a drop-in replacement for the recurrent unit in RomniStereo$_{32}$ [10], as shown in Table 6. While SRU provides a modest improvement over the original ConvGRU, DAA+DiGRU yields substantially larger gains across all three datasets, reducing >1 error on OmniThings by 15.4% and MAE on OmniHouse by 21.3% relative to the original architecture. The consistent improvement demonstrates that DAA+DiGRU provides complementary benefits independent of the underlying pipeline and can serve as a general-purpose upgrade for RAFT-based omnidirectional stereo frameworks.

5 Conclusion

We present PDF-Omni, a geometry-aware omnidirectional stereo matching framework that models spatially varying fisheye distortion through a Poincaré Dual Disk Distortion Field. The PDD distortion field unifies radial distortion magnitude and view-transition ambiguity, which Distortion-Aware Attention and DiGRU translate into adaptive receptive field allocation for high-uncertainty seam regions. The Local Entropy Cumulative loss further stabilizes convergence by adapting supervision tolerance to cost volume entropy in ambiguous regions. Experiments on five datasets demonstrate state-of-the-art accuracy, and the proposed refinement modules generalize as drop-in replacements for existing omnidirectional stereo architectures.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT) (IITP-2026-RS-2024-00439292)

References

1. Ahmad, O., Lecue, F.: Fisheyehdk: Hyperbolic deformable kernel learning for ultra-wide field-of-view image recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 5968–5975 (2022)
2. Cannon, J.W., Floyd, W.J., Kenyon, R., Parry, W.R., et al.: Hyperbolic geometry. Flavors of geometry **31**(59-115), 2 (1997)
3. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5410–5418 (2018)
4. Chen, Z., Lin, C., Nie, L., Liao, K., Zhao, Y.: Unsupervised omnimvs: Efficient omnidirectional depth inference via establishing pseudo-stereo supervision. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10873–10879. IEEE (2023)
5. Chen, Z., Lin, C., Nie, L., Shen, Z., Liao, K., Cao, Y., Zhao, Y.: S-omnimvs: Incorporating sphere geometry into omnidirectional stereo matching. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1495–1503 (2023)
6. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
7. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3273–3282 (2019)
8. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence **30**(2), 328–341 (2007)
9. Im, S., Ha, H., Rameau, F., Jeon, H.G., Choe, G., Kweon, I.S.: All-around depth from small motion with a spherical panoramic camera. In: European Conference on Computer Vision. pp. 156–172. Springer (2016)

10. Jiang, H., Xu, R., Tan, M., Jiang, W.: Romnistereo: Recurrent omnidirectional stereo matching. IEEE Robotics and Automation Letters **9**(3), 2511–2518 (2024)
11. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE international conference on computer vision. pp. 66–75 (2017)
12. Li, J., Wang, P., Xiong, P., Cai, T., Yan, Z., Yang, L., Liu, J., Fan, H., Liu, S.: Practical stereo matching via cascaded recurrent network with adaptive correlation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16263–16272 (2022)
13. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International conference on 3D vision (3DV). pp. 218–227. IEEE (2021)
14. Meuleman, A., Jang, H., Jeon, D.S., Kim, M.H.: Real-time sphere sweeping stereo from multiview fisheye images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11423–11432 (2021)
15. Min, J., Jeon, Y., Kim, J., Choi, M.: S2m2: Scalable stereo matching model for reliable depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26729–26739 (2025)
16. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. Advances in neural information processing systems **30** (2017)
17. Son, E., Jo, H., Kwon, W., Lee, S.J.: Mdp-omni: Parameter-free multimodal depth prior-based sampling for omnidirectional stereo matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26178–26187 (2025)
18. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in neural information processing systems **33**, 7537–7547 (2020)
19. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
20. Wang, P., Li, M., Cao, J., Du, S., Li, Y.: Casomnimvs: Cascade omnidirectional depth estimation with dynamic spherical sweeping. Applied Sciences **14**(2), 517 (2024)
21. Wang, X., Xu, G., Jia, H., Yang, X.: Selective-stereo: Adaptive frequency information selection for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19701–19710 (2024)
22. Wang, Y., Yang, Y., Deng, J., Meng, H., Chen, G.: Fastomnimvs: Real-time omnidirectional depth estimation from multiview fisheye images. In: 2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS). pp. 1311–1318. IEEE (2023)
23. Wei, X., Ran, Z., Lu, X.: Dcpb: Deformable convolution based on the poincare ball for top-view fisheye cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13308–13317 (2023)
24. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Rao, Y., Huang, G., Lu, J., Zhou, J.: Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In: Conference on robot learning. pp. 539–549. PMLR (2023)
25. Won, C., Ryu, J., Lim, J.: Omnimvs: End-to-end learning for omnidirectional stereo matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8987–8996 (2019)

26. Won, C., Ryu, J., Lim, J.: Sweepnet: Wide-baseline omnidirectional depth estimation. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 6073–6079. IEEE (2019)
27. Won, C., Ryu, J., Lim, J.: End-to-end learning for omnidirectional stereo matching with uncertainty prior. IEEE transactions on pattern analysis and machine intelligence **43**(11), 3850–3862 (2020)
28. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
29. Xie, S., Wang, D., Liu, Y.H.: Omnividar: Omnidirectional depth estimation from multi-fisheye images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21529–21538 (2023)
30. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21919–21928 (2023)
31. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 185–194 (2019)