DECOUPLING SPHERICAL REASONING FROM DENSE PREDICTION FOR 360° DEPTH ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The equirectangular projection (ERP) is widely used for panoramic depth estimation, but its spatially varying distortion makes geometry-consistent feature modeling challenging. We revisit panoramic depth estimation by decoupling contextual modeling in native spherical space from dense ERP prediction. To this end, we propose a Fibonacci Spherical Graph (FSG) as an intermediate reasoning space to lift ERP features onto quasi-uniform Fibonacci nodes on the sphere and capture local and long-range dependencies through complementary spherical neighborhoods. The resulting spherical discretization distributes graph nodes approximately uniformly over the spherical surface, reducing the over-representation of highly stretched regions during relational modeling. Operating on a compact set of Fibonacci nodes also avoids the computational burden of constructing and processing a graph at full ERP resolution. To bridge spherical reasoning and dense prediction, we propose a Spherical Context Conditioning (SCC) module that adaptively modulates dense ERP features with the enhanced spherical representation, allowing spherical context to guide pixel-aligned depth prediction. Extensive experiments on three benchmarks demonstrate that the proposed method consistently achieves superior depth accuracy over existing approaches.

1 INTRODUCTION

Panoramic images capture a complete 360° observation of the surrounding environment and provide rich spatial information for scene understanding, robotic navigation, augmented reality, and 3D reconstruction (Chang et al., 2017; Albanis et al., 2021). Monocular panoramic depth estimation aims to recover dense scene depth from a single panoramic observation and serves as a fundamental task for understanding surrounding 3D environments. Unlike perspective images, panoramic images represent a spherical observation of a scene, where spatial relationships are naturally defined on the sphere rather than on a planar image domain.

To process panoramic images with existing vision architectures, most methods adopt the equirectangular projection (ERP), which unfolds the spherical image into a regular planar representation and enables the direct application of mature dense prediction networks (Zioulis et al., 2018; Wang et al., 2020). However, this spherical-to-planar transformation inevitably alters the spatial organization of the scene. Due to the latitude-dependent distortion of ERP, image-space distances and neighborhood structures undergo different degrees of stretching across the panorama (Tateno et al., 2018; Chen et al., 2021). Consequently, pixel neighborhoods defined on the ERP grid do not uniformly reflect spatial relationships on the sphere. A local window of identical size can correspond to different spherical extents and geometric structures at different locations, causing feature aggregation to capture inconsistent scene contexts. As a result, models operating directly on ERP representations may learn contextual dependencies influenced by the projection space rather than the spherical geometry of the scene, making geometry-consistent feature modeling challenging for panoramic depth estimation.

Existing panoramic depth estimation methods mainly alleviate these challenges through projection-aware feature learning. ERP-based approaches adapt convolution operations by modifying sampling locations or receptive fields according to spatially varying distortion in the equirectangular domain (Tateno et al., 2018; Coors et al., 2018; Khasanova & Frossard, 2019). Alternatively, multi-projection approaches introduce additional representations, such as cubemap (Wang et al., 2020;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2022; Shen et al., 2026) or tangent views (Li et al., 2022; Shen et al., 2022), to complement ERP features and reduce the impact of severe projection distortion (Wang et al., 2020; Jiang et al., 2021; Li et al., 2022; Ai et al., 2023; Ai & Wang, 2024). Recent methods further incorporate attention mechanisms, geometric constraints, and cross-projection fusion to enhance contextual modeling (Shen et al., 2022; Ai et al., 2023; Yun et al., 2023; Li et al., 2023). Although these approaches have achieved significant progress, feature interactions in existing panoramic depth estimation pipelines are still mainly performed in equirectangular or other planar projection spaces. Consequently, the contextual relationships learned by these methods remain constrained by the spatial organization of the projection space, rather than being directly established according to spherical geometry.

This observation motivates us to reconsider how feature relationships are modeled in panoramic depth estimation. Rather than continuously adapting planar representations to compensate for projection distortion, we aim to construct a geometry-consistent spherical reasoning space where feature interactions are established according to the geometric structure of the sphere. Since the spherical surface does not naturally follow a regular planar grid (Cohen et al., 2018; Jiang et al., 2019; Lee et al., 2019), such a reasoning space requires a representation that can flexibly describe spherical neighborhoods and relational interactions. This provides more appropriate support for contextual modeling by allowing relationships among features to be aggregated according to their geometric configurations on the sphere. Importantly, the spherical reasoning space is designed to complement rather than replace ERP representations, since ERP remains effective for dense pixel-aligned prediction.

Based on this principle, we introduce a geometry-consistent spherical reasoning space for monocular panoramic depth estimation. Instead of performing all feature interactions in the distorted ERP domain or replacing ERP prediction with spherical decoding, we decouple contextual reasoning from dense prediction. Specifically, we instantiate this reasoning space with a Fibonacci Spherical Graph (FSG), where spherical relationships are modeled independently from the pixel-aligned ERP representation. Instead of defining contextual interactions solely on projected image grids, FSG establishes feature relationships directly over spherical geometry, allowing spatial dependencies to be modeled independently of the latitude-dependent deformation of ERP. Fibonacci discretization provides a compact and approximately uniform support for this reasoning process, making spherical relational modeling practical without replacing the dense ERP representation.

To bridge spherical reasoning and dense prediction, we further introduce a Spherical Context Conditioning (SCC) module. SCC transfers geometry-aware spherical context back to ERP features while preserving the dense spatial correspondence required for depth estimation. Hence, FSG reasons over scene geometry on the sphere, while ERP features remain responsible for spatially aligned depth estimation.

Extensive experiments on three widely used panoramic depth benchmarks demonstrate that the proposed method consistently improves depth estimation accuracy over existing approaches. Ablation studies further verify the effectiveness of spherical relational modeling and spherical context conditioning. Our main contributions are summarized as follows:

- We introduce a new formulation for panoramic depth estimation that separates geometry-aware spherical reasoning from dense ERP prediction.
- We propose a Fibonacci Spherical Graph (FSG) as a compact intermediate reasoning space, where contextual dependencies are modeled according to spherical geometry rather than ERP sampling topology.
- We design a Spherical Context Conditioning (SCC) module to effectively transfer spherical context into pixel-aligned ERP features for dense depth estimation.

2 Related Work

Monocular panoramic depth estimation. Most monocular panoramic depth estimation methods adopt equirectangular projection (ERP) because it preserves a dense rectangular layout and is compatible with standard CNN- and Transformer-based architectures (Zioulis et al., 2018). However, the latitude-dependent distortion of ERP causes inconsistent spatial relationships across the panorama. Early works address this issue through distortion-aware convolutions, adaptive sam-

pling, and geometry-aware receptive fields (Tateno et al., 2018; Chen et al., 2021; Su & Grauman, 2019), but feature interactions are still defined on the distorted ERP grid.

Another line of research combines ERP with complementary projections. BiFuse and UniFuse fuse ERP and cubemap features (Wang et al., 2020; Jiang et al., 2021), while OmniFusion, HRDFuse, and PanoFormer exploit tangent or multi-projection representations to enhance contextual modeling (Li et al., 2022; Ai et al., 2023; Shen et al., 2022). Although these approaches improve distortion handling and global context, feature relationships are still established within individual projection domains or through cross-projection fusion, rather than being explicitly organized according to spherical geometry.

Recent methods further explore spherical representations for panoramic depth estimation. SphereDepth performs feature extraction and prediction on spherical meshes (Yan et al., 2022), $\mathcal{S}^2$Net employs equal-area HEALPix grids for spherical decoding (Li et al., 2023), and Elite360D combines ERP features with an icosahedral point representation through bi-projection fusion (Ai & Wang, 2024). These methods demonstrate the effectiveness of spherical geometry, but the sphere is mainly explored as a prediction domain or an auxiliary representation. In contrast, we introduce a compact spherical reasoning space where feature relationships are modeled according to spherical geometry while dense depth prediction remains in ERP space.

Spherical representation and geometric modeling. Spherical representations have been widely studied for omnidirectional vision. Spherical CNNs and equivariant networks define operators directly on spherical domains (Cohen et al., 2018; Esteves et al., 2018), while SphereNet and related approaches adapt convolution operations to address ERP distortion (Coors et al., 2018; Lee et al., 2019; Jiang et al., 2019). These studies mainly focus on constructing geometry-aware feature representations for spherical signals.

Graph-based spherical models provide a flexible formulation for irregular spherical structures. DeepSphere represents the sphere as a graph and performs graph convolution over spherical neighborhoods (Defferrard et al., 2020). Recent works have also explored Fibonacci-based spherical discretizations for efficient spherical representations, such as organizing spherical Gaussian primitives for high-resolution panorama synthesis (Zhang et al., 2025).

Different from these approaches, FSG employs Fibonacci nodes as a compact spherical reasoning space rather than a final representation or prediction domain. The proposed design introduces spherical geometry-aware relational modeling while retaining ERP for dense pixel-aligned prediction.

3 METHODOLOGY

3.1 OVERVIEW

Panoramic depth estimation requires contextual reasoning across the spherical field of view while preserving the pixel correspondence needed for dense prediction. ERP offers a convenient representation for the latter, but its latitude-dependent distortion makes image-grid neighborhoods an uneven basis for the former. We therefore assign these roles to complementary representations: a Fibonacci Spherical Graph (FSG) organizes contextual interaction on the sphere, and Spherical Context Conditioning (SCC) transfers the resulting context to dense ERP features.

Figure 1(a) shows how this separation is incorporated into UniFuse (Jiang et al., 2021). We retain its dual-projection encoders, fusion modules, and ERP decoder, and independently refine the ERP features at four scales, from $1/4$ to $1/32$ resolution, before their corresponding fusion stages. The FSG branch in Figure 1(b) maps an ERP feature X_l to spherical context G_l, which the SCC module in Figure 1(c) uses to produce the refined feature $\hat{X}_l$. For a feature map $X_l \in \mathbb{R}^{B\times C_l\times H_l\times W_l}$, this process is

$$\begin{aligned} V_l &= \mathcal{S}_l(X_l, \mathcal{P}_l), \quad Z_l = \mathcal{R}_l(V_l; \mathcal{P}_l), \\ G_l &= \mathcal{B}_l(Z_l, \mathcal{P}_l), \quad \hat{X}_l = X_l + \mathcal{C}_l(X_l, G_l). \end{aligned} \tag{1}$$

Here, $\mathcal{S}_l$ samples features onto spherical nodes $\mathcal{P}_l$, $\mathcal{R}_l$ models their relationships, $\mathcal{B}_l$ returns the enhanced context to ERP, and $\mathcal{C}_l$ produces a residual update. The sphere thus serves as an intermediate reasoning space within the dense prediction pipeline. We omit the scale index below.

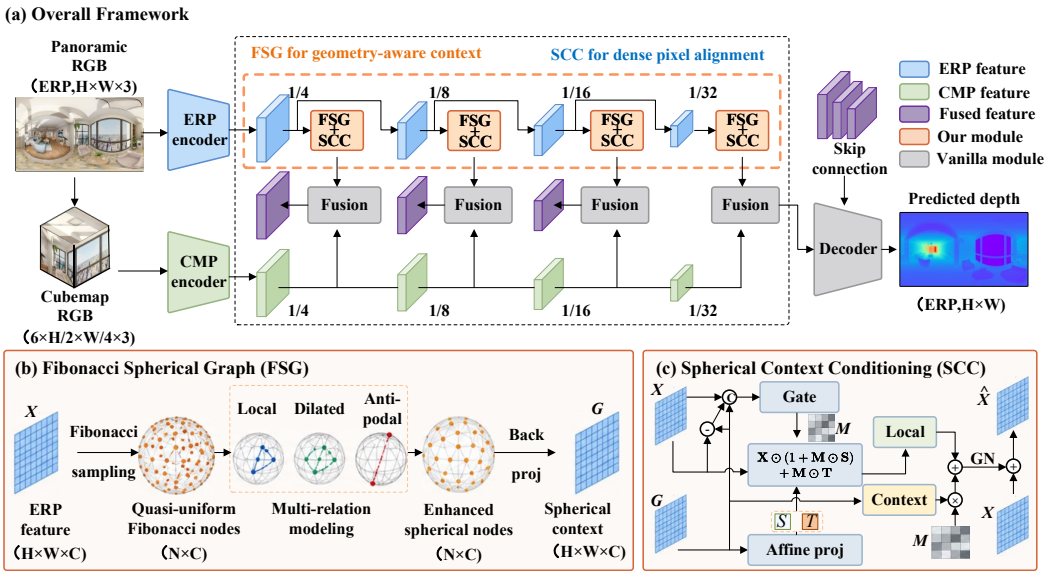

Figure 1: Overview of the proposed framework. (a) FSG and SCC refine multi-scale ERP features before ERP–cubemap fusion. (b) FSG samples features onto Fibonacci nodes, aggregates local, dilated, and antipodal context, and back-projects the enhanced features to ERP. (c) SCC uses spherical context G to condition dense ERP features X through gated modulation and context injection. The schematic is elaborated by the operators and residual update in Sec. 3.4.

3.2 Fibonacci Spherical Representation

The first consideration is how to distribute the support for spherical reasoning. Uniformly spaced ERP pixels represent progressively smaller spherical areas toward the poles. Using these locations as graph nodes therefore allocates a disproportionate share of the node budget to high-latitude regions. We use N Fibonacci nodes $\mathcal{P} = \{\mathbf{p}_i\}_{i=0}^{N-1}$ on the unit sphere, combining uniformly spaced vertical coordinates with golden-angle increments to obtain approximately uniform spherical coverage.

Following the sampling stage in Figure 1(b), each node gathers an ERP feature by bilinear sampling at its corresponding viewing direction, producing $V \in \mathbb{R}^{B\times N\times C}$. Horizontal circular padding connects samples across the ERP seam. This construction gives relational modeling a more balanced distribution of viewing directions while allowing its node count to be chosen independently of ERP resolution. The original dense features remain available for subsequent prediction, and gradients propagate through sampling to the encoder. Node coordinates are fixed; their construction and sampling mappings are provided in Appendix A.1.

3.3 Multi-Relation Spherical Graph Reasoning

A balanced node distribution provides the support for reasoning, but the connections determine which evidence can be brought together. Nearby viewing directions provide local context, while larger scene structures extend across broader angular ranges. We therefore construct the local, dilated, and antipodal relations illustrated in Figure 1(b) over the same nodes. The local relation selects the k nearest nodes, including the center node. The dilated relation selects the next k nodes in the distance ordering, extending interaction beyond the immediate neighborhood. The antipodal relation selects the k nodes nearest to $-\mathbf{p}_i$, providing direct communication with the opposite part of the panorama.

For relation r, let $\mathcal{N}_r(i)$ denote the neighbors of node i. Given input node features h_i, relation-specific aggregation is

$$h_i^r = \phi_r\left(\sum_{j\in\mathcal{N}_r(i)} \alpha_{ij}^r \left(W_r h_j + E_r \mathbf{g}_{ij}\right)\right), \tag{2}$$

where W_r and E_r project features and relative geometry, and $\mathbf{g}_{ij}$ encodes node displacement and angular separation. The operator ϕ_r applies the post-aggregation transformation. Affinities α_{ij}^r are fixed softmax-normalized cosine similarities, measured from $\mathbf{p}_i$ for local and dilated relations and from $-\mathbf{p}_i$ for the antipodal relation. To let the contribution of each angular range depend on the observed features, we combine the relation responses using node-wise weights:

$$\beta_i^r = \mathrm{softmax}_r\big(\mathbf{a}_r^\top h_i^r + b_r\big), \qquad u_i = \sum_r \beta_i^r h_i^r. \tag{3}$$

Spherical geometry determines which directions exchange information and how messages are aggregated within each relation; learned relation weights adapt their combination at each node. This allows nearby evidence and distant context to contribute differently across the panorama. Residual graph blocks with feed-forward refinement produce the enhanced representation Z. Complete weight and block definitions are given in Appendix A.2.

3.4 Spherical Context Conditioning

The remaining challenge is to make spherical reasoning useful for dense prediction without making pixel-level detail depend solely on the graph representation. The back-projection stage in Figure 1(b) interpolates enhanced node features according to each ERP pixel's viewing direction, obtaining a context map $G \in \mathbb{R}^{B\times C\times H\times W}$. This restores spatial correspondence, but interpolation aggregates node features and does not guarantee recovery of the original dense feature variations. SCC, illustrated in Figure 1(c), therefore keeps the ERP feature X as the carrier of dense information and uses G to condition its update. The interpolation rule is provided in Appendix A.3.

The affine projection in Figure 1(c) predicts a bounded scale adjustment S and an additive shift T from spherical context. The Gate branch jointly uses the ERP feature, spherical context, and their difference to control modulation at each spatial location and channel. Making the normalization operations explicit, let $\bar{X}$ and $\bar{G}$ denote separately group-normalized features. The conditioning update is

$$\begin{aligned} [S_0, T] &= f_{\text{aff}}(\bar{G}), & S &= a\tanh(S_0), \\ M &= \sigma\big(f_g([\bar{X}, \bar{G}, \bar{X} - \bar{G}])\big), \\ X_c &= X \odot (1 + M \odot S) + M \odot T, \\ \mathcal{C}(X, G) &= \mathcal{N}_o\big(f_{\text{local}}(X_c) + M \odot f_{\text{ctx}}(\bar{G})\big). \end{aligned} \tag{4}$$

Here, $a = 0.5$, σ is sigmoid, brackets denote channel concatenation, and $\odot$ denotes element-wise multiplication. The Local branch processes X_c through a depth-wise convolution with horizontal circular padding. The Context branch projects $\bar{G}$, and its response is multiplied by the same gate M. GroupNorm $\mathcal{N}_o$ normalizes the combined update, which is added to the original X as specified in Eq. 1. This coupling lets spherical relationships influence both the transformation of existing ERP features and the context added to them, while retaining a direct dense feature pathway. Operator details and module configurations are given in Appendices A.4 and A.5.

3.5 Objective Function

Following previous works (Shen et al., 2026; Lee et al., 2025), our objective function consists of two terms. We use the BerHu loss $\mathcal{L}_{\text{BerHu}}$ for depth regression. Given the predicted depth $\hat{D}$, ground-truth depth D, and the set of valid pixels Ω_Q, it is defined as

$$\mathcal{L}_{\text{BerHu}} = \frac{1}{|\Omega_Q|} \sum_{p\in\Omega_Q} \begin{cases} e_p, & e_p \le c, \\ \dfrac{e_p^2 + c^2}{2c}, & e_p > c, \end{cases} \qquad e_p = |\hat{D}_p - D_p|, \tag{5}$$

where $c = 0.2$ denotes the BerHu transition threshold.

We further adopt the spherical gradient loss $\mathcal{L}_{\text{sg}}$ introduced in PGFuse (Shen et al., 2026) to supervise spatial depth variations,

$$\mathcal{L}_{\text{sg}} = \frac{1}{|\Omega_Q^\nabla|} \sum_{p\in\Omega_Q^\nabla} \left\| \nabla_s \hat{D}_p - \nabla_s D_p \right\|_1, \tag{6}$$

Table 1: Quantitative comparison with existing panoramic depth estimation methods. For fair comparison, * denotes re-evaluation under the Elite360D protocol, while † denotes retraining with a ResNet-34 backbone. All remaining baseline results are taken from PGFuse (Shen et al., 2026). Best and second-best results are shown in bold and underlined, respectively.

Dataset	Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$\delta_1(\%)$ ↑	$\delta_2(\%)$ ↑	$\delta_3(\%)$ ↑
M3D	EGFormer	0.1473	0.1517	0.6025	81.58	93.90	97.35
	PanoFormer	0.1051	0.0966	0.4929	89.08	96.23	98.31
	BiFuse	0.1126	0.0992	0.5027	88.00	96.13	98.47
	BiFuse++	0.1123	0.0915	0.4853	88.12	96.56	98.69
	UniFuse	0.1144	0.0936	0.4835	87.85	96.59	98.73
	OmniFusion	0.1161	0.1007	0.4931	87.72	96.15	98.44
	HRDFuse	0.1172	0.0971	0.5025	86.74	96.17	98.49
	Elite360D	0.1115	0.0914	0.4875	88.15	96.46	<u>98.74</u>
	PGFuse	0.1067	0.0858	0.4624	89.06	96.75	98.72
	HUSH†	<u>0.0957</u>	<u>0.0798</u>	<u>0.4556</u>	<u>90.91</u>	<u>96.83</u>	98.66
	Ours	**0.0949**	**0.0776**	**0.4455**	**91.16**	**97.01**	**98.87**
S2D3D	EGFormer	0.1528	0.1408	0.4974	81.85	93.38	97.36
	PanoFormer	0.1122	0.0786	0.3945	88.74	95.84	98.59
	OmniFusion	0.1154	0.0775	0.3809	86.74	96.03	98.71
	UniFuse	0.1124	0.0709	0.3555	87.06	97.04	98.99
	Elite360D	0.1182	0.0728	0.3756	88.72	96.84	98.92
	PGFuse*	0.1035	0.0645	0.3680	88.98	96.79	99.00
	HUSH†	<u>0.0982</u>	<u>0.0587</u>	<u>0.3446</u>	<u>89.15</u>	<u>97.46</u>	<u>99.12</u>
	Ours	**0.0949**	**0.0568**	**0.3420**	**89.53**	**97.68**	**99.18**
S3D	EGFormer	0.2205	0.4509	0.6841	79.79	90.71	94.55
	PanoFormer	0.2549	0.4949	0.7937	74.70	89.15	93.97
	BiFuse	0.1573	0.2455	0.5213	85.91	94.00	96.72
	UniFuse	0.1506	0.2319	0.5016	85.42	93.99	96.76
	Elite360D	0.1480	<u>0.2215</u>	0.4961	87.41	94.34	96.66
	PGFuse	<u>0.1401</u>	0.2499	<u>0.4394</u>	87.78	94.49	<u>96.82</u>
	HUSH†	0.1427	0.2489	0.4571	<u>88.45</u>	<u>94.56</u>	96.71
	Ours	**0.1289**	**0.2129**	**0.4046**	**89.12**	**95.25**	**97.09**

where ∇_s denotes the spherical gradient operator and Ω_Q^{∇} contains valid gradient pairs.

The overall objective is formulated as

$$\mathcal{L} = \mathcal{L}_{\mathrm{BerHu}} + \lambda_{\mathrm{sg}} \mathcal{L}_{\mathrm{sg}}, \tag{7}$$

where λ_{sg} is set to 0.5 following PGFuse (Shen et al., 2026) and HUSH (Lee et al., 2025).

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on Matterport3D (Chang et al., 2017), Stanford2D3D (Armeni et al., 2017), and Structured3D (Zheng et al., 2020). The first two datasets contain real indoor panoramas, while Structured3D provides synthetic indoor scenes with depth annotations. Together, they allow us to assess the method on both captured and rendered panoramic observations. We follow the training and evaluation protocols adopted by previous methods (Jiang et al., 2021; Ai & Wang, 2024) for each benchmark.

Evaluation metrics. We report Absolute Relative Error (Abs Rel), Squared Relative Error (Sq Rel), Root Mean Squared Error (RMSE), and threshold accuracies δ_t, $t \in \{1, 2, 3\}$, over valid depth pixels. The latter measure the percentage of pixels satisfying $\max(\hat{D}/D, D/\hat{D}) < 1.25^t$. RMSE emphasizes larger absolute depth errors, whereas the threshold accuracies measure how frequently predictions meet specified relative-error bounds.

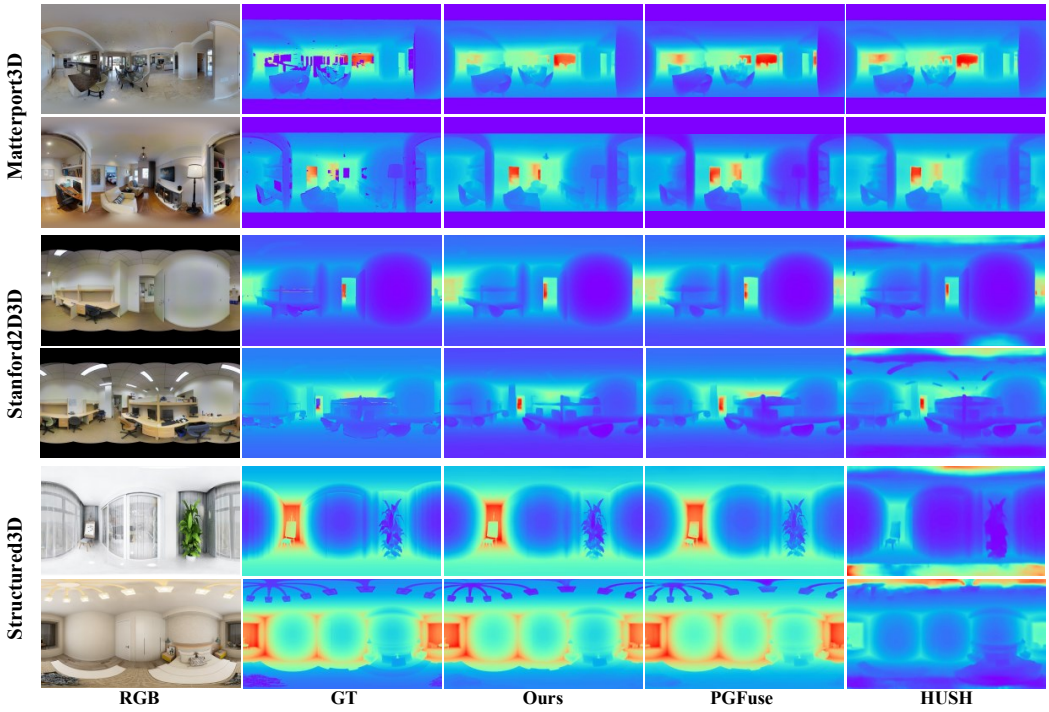

Figure 2: Qualitative comparison results on Matterport3D, Stanford2D3D, and Structured3D. Best viewed in color.

Implementation details. We implement our approach using the PyTorch framework and conduct all experiments on a single NVIDIA GeForce RTX 3090 GPU. We adopt a ResNet-34 backbone initialized with ImageNet-1K pretrained weights. The model is trained using the Adam optimizer with an initial learning rate of 1×10^{-4}, and a batch size of 8. Other optimizer hyperparameters follow the default settings. During training, we employ horizontal flipping, horizontal circular rotation, and luminance augmentation following previous panoramic depth estimation methods (Wang et al., 2020; Jiang et al., 2021). The node counts, neighborhood sizes, and graph depths are specified in Appendix A.5; sampling, propagation, and conditioning details are given in Appendices A.1–A.4.

4.2 Comparison with State-of-the-Art Methods

Quantitative comparison. Table 1 compares our method with existing panoramic depth estimation approaches, including EGFormer (Yun et al., 2023), PanoFormer (Shen et al., 2022), BiFuse (Wang et al., 2020), BiFuse++ (Wang et al., 2022), UniFuse (Jiang et al., 2021), OmniFusion (Li et al., 2022), HRDFuse (Ai et al., 2023), Elite360D (Ai & Wang, 2024), PGFuse (Shen et al., 2026), and HUSH (Lee et al., 2025). Our method achieves the lowest errors and highest threshold accuracies among the compared methods on all three datasets.

Table 1 shows that our method consistently outperforms existing approaches on all three benchmarks, with strong performance in both RMSE and δ metrics. This indicates that the method reduces large depth errors while maintaining reliable prediction across the panorama. The improvement is closely related to our decoupled design, where FSG performs context modeling in the native spherical space to alleviate ERP distortion and sampling imbalance, while SCC transfers the resulting spherical context back to the dense ERP stream for pixel-aligned depth prediction.

Qualitative comparison. Figure 2 presents qualitative comparisons on all three benchmarks. On Matterport3D, our predictions follow the ground-truth depth variation more consistently over extended walls and floor regions while preserving clear transitions around openings and furniture. On Stanford2D3D, large surfaces exhibit more coherent depth variation, and distant structures re-

Table 2: Ablation of node support, graph reasoning, and SCC on Matterport3D. ERP Grid uses uniformly sampled ERP nodes with matched node and edge counts.

Support	Graph	SCC	RMSE ↓	Abs Rel ↓	δ_1 ↑
ERP	–	–	0.4737	0.1036	88.36
Fibonacci	✓	–	0.4617	0.0968	90.62
Fibonacci	–	✓	0.4577	0.1017	90.01
ERP Grid	✓	✓	0.4558	0.0999	90.29
Fibonacci	✓	✓	**0.4455**	**0.0949**	**91.16**

main better aligned with the ground truth. Similar behavior is observed on Structured3D, where our method better preserves room-scale geometry and the spatial organization of extended surfaces.

Local depth transitions around object and structural boundaries are also well preserved rather than being over-smoothed by broader contextual aggregation. These visual results complement the quantitative evaluation by showing that the proposed method improves large-scale structural consistency while retaining local geometric detail.

4.3 Ablation Studies and Analysis

We conduct ablations on Matterport3D to examine the contributions of graph reasoning, SCC, and spherical node support. Table 2 compares five controlled variants. The first row corresponds to the UniFuse baseline retrained with the spherical gradient loss. To isolate graph reasoning without SCC, we back-project the graph-enhanced Fibonacci features to the ERP space and add them element-wise to the original ERP features (the second row). Then the sampled Fibonacci features are directly back-projected without graph reasoning and incorporated into the ERP stream through SCC (the third row). We then replace Fibonacci nodes with uniformly sampled ERP-grid nodes while retaining graph reasoning and SCC (the fourth row). The complete model combines Fibonacci-based graph reasoning with SCC (the fifth row). The ERP-grid and Fibonacci variants use matched node and edge counts, allowing the influence of node support to be examined independently.

Effect of spherical node support. From Table 2, we can observe that replacing ERP-grid nodes with Fibonacci nodes reduces RMSE, while improving δ_1. Since the remaining components are unchanged, this comparison directly reflects the influence of node support. Fibonacci nodes distribute the representation budget more uniformly over the native spherical space, reducing the latitude-dependent redundancy introduced by ERP-grid sampling.

Effect of graph reasoning. Adding graph reasoning to the Fibonacci representation improves both RMSE and δ_1. Fibonacci sampling determines how viewing directions are distributed over the sphere, whereas graph reasoning explicitly establishes dependencies among them. The additional gain shows that a balanced spherical support alone is insufficient and that contextual interaction over the spherical topology contributes further to depth estimation.

Effect of SCC. As shown in Table 2, SCC consistently improves all evaluation metrics over direct residual fusion. This suggests that simply injecting the back-projected spherical features into the ERP stream is insufficient. By adaptively conditioning the dense ERP features with graph-derived spherical context, SCC enables more selective information transfer and better preserves spatially localized depth structures.

Table 3: Ablation of spherical relations on Matterport3D. Fibonacci support and SCC are retained.

Relations	RMSE ↓	Abs Rel ↓	δ_1 ↑
Local	0.4577	0.0994	90.42
Local + Dilated	0.4615	0.1006	90.34
Local + Antipodal	0.4501	0.0957	90.96
Local + Dilated + Antipodal	**0.4455**	**0.0949**	**91.16**

Effect of spherical relation configurations. Table 3 shows that adding antipodal connections clearly improves the local graph, whereas adding the dilated relation to the local graph does not, indicating that simply enlarging the interaction range is insufficient. The best performance is achieved by combining all three relations, suggesting that local, intermediate-range, and long-range interactions provide complementary structural context. Adaptive relation fusion further integrates these context paths for effective multi-range spherical reasoning.

Effect across latitude regions. Table 4 further compares ERP-grid and Fibonacci support across different latitude regions. Fibonacci support improves all three regions, with the largest gain observed in the equatorial region. Under uniform ERP-grid sampling, a fixed node budget is distributed uniformly in image coordinates even though equal latitude intervals correspond to unequal spherical surface areas. This results in relatively redundant support toward high latitudes and comparatively insufficient representation of the lower-distortion equatorial region. Redistributing the nodes more uniformly over the sphere alleviates this imbalance and assigns more effective representation capacity to regions with larger spherical coverage and less projection distortion. The more pronounced improvement near the equator is consistent with this motivation, while the gains across all latitude regions indicate that the benefit is not confined to a particular part of the panorama.

Table 4: Latitude-wise comparison of ERP-grid and Fibonacci node support on Matterport3D. Both variants retain the same graph reasoning and SCC configuration.

Support	Region	RMSE ↓	Abs Rel ↓	δ_1 ↑
ERP Grid	Equatorial (30°S–30°N)	0.5736	0.1151	87.60
Fibonacci	Equatorial (30°S–30°N)	**0.5585**	**0.1073**	**88.99**
ERP Grid	Mid-latitude (30°S–60°S, 30°N–60°N)	0.3219	0.0894	92.19
Fibonacci	Mid-latitude (30°S–60°S, 30°N–60°N)	**0.3151**	**0.0861**	**92.70**
ERP Grid	High-latitude (60°S–90°S, 60°N–90°N)	0.2676	0.0822	93.29
Fibonacci	High-latitude (60°S–90°S, 60°N–90°N)	**0.2605**	**0.0817**	**93.48**

Computational complexity analysis. Table 5 shows that our method introduces moderate computational overhead over UniFuse. However, compared with PGFuse and HUSH, which achieve comparable depth estimation accuracy, our method requires fewer FLOPs and less training memory and achieves lower inference latency. Spherical reasoning is performed on a compact set of Fibonacci nodes rather than full-resolution ERP features, enabling low-cost contextual modeling.

Table 5: Computational comparison at an input resolution of 512×1024 with a batch size of 1, where FLOPs are computed using `fvcore.nn.FlopCountAnalysis`.

Model	FLOPs ↓	Train Mem. (GB) ↓	Inference latency ↓
UniFuse	96G	2.96	21ms
PGFuse	126G	3.63	41ms
HUSH	237G	5.11	47ms
Ours	104G	3.37	38ms

5 Conclusion

We introduced a Fibonacci Spherical Graph as an intermediate reasoning space for monocular panoramic depth estimation. Quasi-uniform spherical nodes and multiple geometric relations organize contextual interaction, while SCC incorporates the resulting context into dense ERP features. This separates spherical relational modeling from pixel-aligned prediction within the existing dual-projection pipeline. Experiments on Matterport3D, Stanford2D3D, and Structured3D demonstrate improved accuracy over the compared methods. The controlled node-support comparison and component ablations support the benefits of spherical sampling and the joint use of graph reasoning and conditioning. These findings suggest that choosing a dedicated representation for contextual interaction is a promising strategy for incorporating panoramic geometry into dense prediction pipelines.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and stylistic refinement. The experimental analyses, numerical results, and scientific conclusions were produced by the authors.

Additionally, generative AI tools were used to improve the presentation and clarity of author-written technical descriptions. All AI-assisted text was manually reviewed and revised to ensure consistency with the actual method, implementation, and experimental results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on monocular panoramic depth estimation and does not involve human-subject studies, collection of personal data, or the release of new datasets. All experiments are conducted on existing research datasets following their established evaluation protocols. We are not aware of specific ethical risks introduced by the proposed methodology beyond those generally associated with computer vision systems and their downstream deployment.

REPRODUCIBILITY STATEMENT

We provide the methodological details required to reproduce the proposed approach in the main paper, including the Fibonacci spherical representation, multi-relation graph construction, spherical context conditioning, and training objective. The experimental section specifies the datasets, evaluation metrics, implementation settings, and ablation configurations used in our evaluation. Additional architectural and implementation details are provided in the supplementary material. We will release the source code to facilitate reproduction of the reported results.

REFERENCES

Hao Ai and Lin Wang. Elite360d: Towards efficient 360 depth estimation via semantic- and distance-aware bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9926–9935, June 2024.

Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, and Lin Wang. Hrdfuse: Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13273–13282, 2023.

Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiros Sterzentsenko, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. Pano3d: A holistic benchmark and a solid baseline for 360deg depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3722–3732, 2021.

Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

Hong-Xiang Chen, Kunhong Li, Zhiheng Fu, Mengyi Liu, Zonghao Chen, and Yulan Guo. Distortion-aware monocular depth estimation for omnidirectional images. *IEEE Signal Processing Letters*, 28:334–338, 2021.

Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018.

Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 525–541. Springer, 2018.

Michaël Defferrard, Martino Milani, Frédérick Gusset, and Nathanaël Perraudin. Deepsphere: a graph-based spherical cnn. *arXiv preprint arXiv:2012.15000*, 2020.

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 52–68, 2018.

Chiyu Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al. Spherical cnns on unstructured grids. *arXiv preprint arXiv:1901.02039*, 2019.

Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters*, 6(2):1519–1526, 2021.

Renata Khasanova and Pascal Frossard. Geometry aware convolutional filters for omnidirectional images representation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3351–3359. PMLR, 2019.

Jongsung Lee, Harin Park, Byeong-Uk Lee, and Kyungdon Joo. Hush: Holistic panoramic 3d scene understanding using spherical harmonics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16599–16608. IEEE, 2025.

Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360° images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9173–9181. IEEE, 2019.

Meng Li, Senbo Wang, Weihao Yuan, Weichao Shen, Zhe Sheng, and Zilong Dong. $\mathcal{S}^2$net: Accurate panorama depth estimation on spherical surface. *IEEE Robotics and Automation Letters*, 8:1053–1060, 2023.

Yu-yang Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2791–2800, 2022.

Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: panorama transformer for indoor 360 depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 195–211. Springer, 2022.

Zhijie Shen, Chunyu Lin, Lang Nie, Kang Liao, Weisi Lin, and Yao Zhao. Revisiting 360 depth estimation with panogabor: A new fusion perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5979–5991, 2026.

Yu-Chuan Su and Kristen Grauman. Kernel transformer networks for compact spherical convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9434–9443. IEEE, 2019.

Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 707–722, 2018.

Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 462–471, 2020.

Fu-En Wang, Yu-Hsuan Yeh, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. Bifuse++: Self-supervised and efficient bi-projection fusion for 360° depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:5448–5460, 2022.

Qingsong Yan, Qiang Wang, Kaiyong Zhao, Bo Li, Xiaoweo Chu, and Fei Deng. Spheredepth: Panorama depth estimation from spherical domain. In *Proceedings of the International Conference on 3D Vision (3DV)*, pp. 1–10. IEEE, 2022.

Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae Eun Rhee. Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6078–6089. IEEE, 2023.

Cheng Zhang, Haofei Xu, Qianyi Wu, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. Pansplat: 4k panorama synthesis with feed-forward gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11437–11447. IEEE, 2025.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 519–535. Springer, 2020.

Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 453–471. Springer, 2018.