

Article

Distortion-Aware Bi-Projection Fusion for 360° Monocular Depth Estimation via Coordinate Attention

Lichuan Geng¹, Li Ma^{2,*}, Yongzhi Qin², Chenyang He³  and Peng Sun³ ¹ School of Geography and Spatial Information, Xuchang University, Xuchang 461000, China; 12014258@xcu.edu.cn² National Key Laboratory of Intelligent Spatial Information Processing, Beijing 100042, China; cyxhn@126.com³ School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China; mo2n0221@gs.zzu.edu.cn (C.H.); iepengsun@zzu.edu.cn (P.S.)

* Correspondence: marycatbj@163.com

Abstract

Monocular depth estimation for 360° omnidirectional images is essential for immersive scene understanding and 3D reconstruction but remains challenging due to the non-uniform geometric distortions introduced by equirectangular projection (ERP). In particular, ERP suffers from latitude-dependent sampling bias, where the effective receptive field of standard convolution varies with the spherical latitude, leading to inconsistent feature representation and degraded depth prediction in high-distortion regions. To address this problem, this paper proposes a distortion-aware bi-projection fusion framework that integrates the global contextual continuity of ERP with the locally distortion-reduced geometric representation of cube map projection (CMP). The core component of the proposed framework is a Multi-scale Coordinate-Transformer Fusion (MCTF) module, which combines convolutional feature mixing, Transformer-based global context modeling, and Coordinate Attention-based spatial recalibration. By explicitly encoding vertical coordinate information into the fusion process, MCTF adaptively recalibrates feature responses according to latitude-dependent distortion patterns. Extensive experiments on the 3D60, Matterport3D, and Stanford2D3D benchmarks demonstrate that the proposed method consistently outperforms state-of-the-art omnidirectional depth estimation approaches. On the 3D60 dataset, our method reduces RMSE by 21.2% compared with the ERP baseline and achieves a 30% RMSE reduction in polar regions, where ERP distortion is most severe. These results validate the effectiveness of coordinate-aware feature calibration for robust 360° monocular depth estimation.

Keywords: 360° vision; omnidirectional depth estimation; monocular depth estimation; bi-projection fusion; Coordinate Attention; geometric distortion



Academic Editor: John A. Kalomiros

Received: 8 May 2026

Revised: 2 June 2026

Accepted: 17 June 2026

Published: 13 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Omnidirectional depth estimation is a fundamental task for immersive 3D scene understanding, with broad applications in autonomous driving, robotic navigation, virtual reality, and indoor reconstruction. Compared with perspective images, 360° omnidirectional images provide a complete spherical field of view and preserve holistic spatial context in a single capture, making them valuable for full-scene perception and 3D reconstruction [1]. Equirectangular projection (ERP) is the most widely used representation because it maps spherical images onto a regular 2D grid and is compatible with standard convolutional

neural networks. However, ERP inevitably introduces severe non-uniform geometric distortion [2].

The main challenge of ERP is its latitude-dependent sampling bias, as illustrated in Figure 1. When a spherical surface is mapped to a planar grid, pixels at different latitudes correspond to unequal areas on the sphere. As the latitude approaches the poles, the horizontal sampling density increases, causing the effective receptive field of a standard convolution kernel to vary approximately with $1/\cos\theta$, where θ denotes the spherical latitude. This violates the spatial stationarity assumption of conventional CNNs and leads to inconsistent feature extraction, especially in high-distortion polar regions.

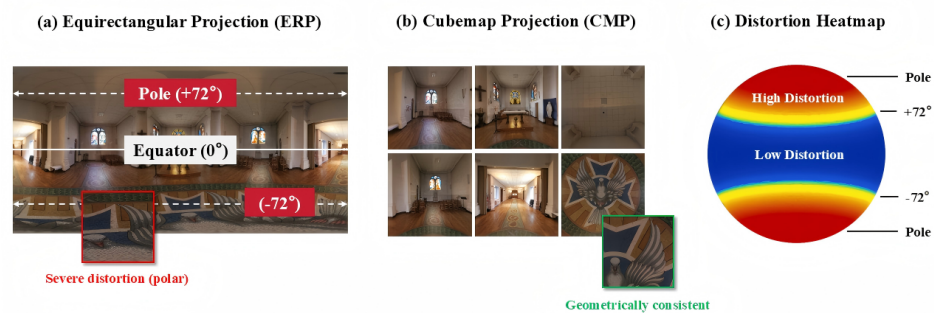


Figure 1. Visualization of ERP distortion and latitude-dependent sampling density. As the latitude θ increases toward the poles, the horizontal sampling density becomes progressively higher, resulting in non-uniform effective receptive fields for standard convolution operations.

Existing methods address this problem mainly through distortion-aware operations or alternative projections. Spherical convolutions and adaptive kernels explicitly model spherical geometry but often introduce additional computational cost and implementation complexity [2]. Cube map projection (CMP) reduces local distortion by projecting the sphere onto six perspective faces, but it may introduce discontinuities across cube boundaries and weaken global continuity. Recent bi-projection methods, such as BiFuse and UniFuse, combine ERP and CMP to exploit their complementary advantages [3,4]. Nevertheless, most existing fusion strategies treat spatial regions uniformly and do not explicitly account for the latitude-dependent distortion pattern of ERP.

To address this limitation, we propose a distortion-aware bi-projection fusion framework for 360° monocular depth estimation. The framework jointly exploits ERP for global contextual continuity and CMP for locally distortion-reduced geometric representation. To fuse these two projection domains effectively, we introduce a Multi-scale Coordinate-Transformer Fusion (MCTF) module, which combines convolutional feature mixing, Transformer-based global context modeling, and Coordinate Attention-based spatial recalibration. Unlike generic attention mechanisms, Coordinate Attention is used here as a latitude-aware calibration mechanism, since the vertical coordinate in ERP directly corresponds to spherical latitude [5].

The main contributions of this paper are summarized as follows:

- We propose a distortion-aware bi-projection fusion framework for 360° monocular depth estimation, combining the global continuity of ERP with the locally distortion-reduced representation of CMP.
- We introduce the Multi-scale Coordinate-Transformer Fusion (MCTF) module, which integrates convolutional mixing, Transformer-based global context modeling, and Coordinate Attention-based spatial recalibration.
- We establish a connection between ERP latitude-dependent distortion and coordinate-aware feature calibration, showing that the vertical encoding in Coordinate Attention can help compensate for spatially varying distortion.

- We conduct extensive experiments on 3D60, Matterport3D, and Stanford2D3D, including standard benchmark comparisons and fine-grained latitude-wise evaluation to verify robustness in high-distortion polar regions.

The remainder of this paper is organized as follows. Section 2 reviews monocular depth estimation, panoramic depth estimation, multi-projection fusion, and attention mechanisms. Section 3 presents the overall architecture and the MCTF module. Section 4 reports quantitative comparisons, ablation studies, latitude-wise analysis, complexity analysis, and qualitative results. Section 5 summarizes the paper and discusses future research directions.

2. Related Work

The development of monocular depth estimation has fundamentally transitioned from regular perspective images to omnidirectional panoramas. Early deep learning approaches primarily focused on regular fields-of-view, establishing baseline architectures for dense spatial regression. However, directly applying these perspective models to 360° data introduces non-trivial challenges due to severe geometric distortions. To mitigate this boundary and sampling variance, recent research has branched into multi-projection fusion strategies, which seek a synergy between different geometric representations. In this section, we review the technical trajectory of these methodologies and contextualize the specific focus of our work.

2.1. Monocular Depth Estimation

Monocular depth estimation aims to infer dense depth information from a single RGB image. Since a single image does not provide explicit geometric scale, this task is inherently ill-posed and requires strong scene priors. Early methods relied on handcrafted visual features and probabilistic graphical models to estimate scene structure, but their generalization ability was limited in complex environments [6]. With the development of deep learning, convolutional neural networks significantly improved monocular depth estimation by learning hierarchical visual representations from large-scale annotated datasets [7–9]. These supervised methods demonstrated that deep networks can effectively capture semantic and structural cues for dense depth prediction.

To reduce the reliance on ground-truth depth annotations, self-supervised monocular depth estimation methods have been widely studied. These approaches typically exploit photometric consistency between stereo image pairs or adjacent video frames to provide supervisory signals [10,11]. More recently, Transformer-based architectures have been introduced into dense prediction tasks to capture long-range dependencies and improve global scene understanding [12–14]. By modeling global interactions, these methods achieve improved structural consistency compared with purely convolutional networks.

Furthermore, within the broader landscape of 3D scene perception, monocular depth estimation serves as a fundamental pixel-level precursor for higher-level multi-modal tasks. A prominent example is Monocular 3D Visual Grounding (Mono3DVG) [15], which aims to localize 3D bounding boxes of specific objects based on natural language queries from a single perspective image. Both tasks inherently grapple with the fundamental ambiguity of lifting 2D representations into 3D spaces without stereo cues. While Mono3DVG models rely on cross-modal attention to align linguistic and visual concepts at the object level, our panoramic framework operates at a denser, geometric level to resolve omnidirectional distortions. Advancements in capturing distortion-free panoramic depth layouts can potentially offer robust spatial priors to enhance object-level localization in complex, omnidirectional 3D grounding scenarios.

Despite these advances, most existing monocular depth estimation methods are primarily designed for perspective images. They usually assume a pinhole camera model and relatively uniform image sampling. Such assumptions do not hold for 360° omnidirectional images, where spherical-to-planar projection introduces severe non-uniform distortion. Therefore, directly applying perspective depth estimation models to ERP images often leads to inconsistent geometric representation and degraded prediction accuracy, especially near the polar regions.

2.2. Panoramic Depth Estimation

Panoramic depth estimation extends monocular depth estimation to 360° omnidirectional imagery. Unlike perspective images, panoramic images provide a complete spherical field of view, which is beneficial for holistic scene understanding and 3D reconstruction. However, the spherical nature of omnidirectional images also introduces unique geometric challenges. The most commonly used representation, equirectangular projection (ERP), maps a sphere onto a rectangular grid, but suffers from latitude-dependent distortion. Pixels near the poles correspond to non-uniform areas on the sphere, causing spatially inconsistent sampling and feature extraction [1,16].

Existing panoramic depth estimation methods can be broadly divided into two categories. The first category focuses on distortion-aware operations. Spherical convolutions, deformable or adaptive kernels, and distortion-aware CNNs have been proposed to better align convolutional operations with spherical geometry [2,17,18]. These methods explicitly consider the geometric structure of omnidirectional images and can reduce the mismatch between planar convolutions and spherical signals. However, they often involve higher computational cost, specialized implementations, or limited compatibility with standard CNN backbones.

The second category explores alternative projection representations. Cube map projection (CMP) projects the sphere onto six perspective faces, thereby reducing local geometric distortion compared with ERP. Several methods leverage CMP or other projection formats to improve local structural modeling in panoramic depth estimation [19,20]. Nevertheless, CMP introduces discontinuities across cube-face boundaries and may weaken global contextual continuity. As a result, models based solely on CMP may struggle to preserve coherent global structures across the entire 360° field of view.

Although existing methods have improved panoramic depth estimation, many of them are evaluated mainly using global metrics averaged over the entire image. Such evaluation protocols may obscure the performance degradation in high-distortion regions. Since ERP distortion varies systematically with latitude, it is important to analyze model robustness across different latitude bands rather than relying only on global average errors.

2.3. Multi-Projection Panoramic Fusion Frameworks

To simultaneously exploit the continuous field-of-view of ERP and the distortion-free local characteristics of CMP, multi-projection fusion frameworks have emerged as a dominant paradigm in recent years. This lineage of research pioneered a structural shift from separate processing to interactive geometric alignment. BiFuse [3] first introduced a dual-branch network that exchanges features between ERP and CMP via a specialized spherical-surface wrapping module. To optimize the computational overhead of bidirectional passing, UniFuse [4] proposed a unidirectional fusion scheme that systematically transfers less-distorted CMP features to enrich the ERP branch. More recently, this paradigm has evolved toward managing multi-scale features and long-range semantic preservation, exemplified by high-resolution representation frameworks like HRDFuse [21] and patch-wise cross-attention networks like OmniFusion [22].

Despite the progressive success of these fusion frameworks, a critical limitation persists: they generally implement spatial feature fusion in a uniform or location-agnostic manner. Because the geometric stretching in ERP is deterministically anchored to the vertical coordinate (spherical latitude), treating different spatial zones identically inevitably introduces representation ambiguity, particularly in highly distorted polar zones.

While general-purpose attention mechanisms like SE [23] and CBAM [24] have been incorporated in various dense prediction tasks to capture channel or isotropic spatial saliency, they inherently lack the capability to encode structural coordinate variations. Coordinate Attention [5] offers a viable remedy by embedding precise directional and positional information into channel interactions. Distinct from conventional practices that employ such modules as generic feature amplifiers, our work reinterprets and re-engineers coordinate encoding as a deterministic, latitude-aware calibration instrument specifically tailored for panoramic multi-projection fusion, thereby resolving the long-standing trade-off between global structural continuity and localized geometric fidelity.

3. Proposed Method

3.1. Preliminary: Mathematical Formulation of Projections

To establish a rigorous geometric foundation, we formally define the two primary projection formats utilized in this framework: Equirectangular Projection (ERP) and Cube Map Projection (CMP).

Equirectangular Projection (ERP): ERP maps spherical coordinates, latitude $\phi \in [-\pi/2, \pi/2]$ and longitude $\theta \in [-\pi, \pi]$, onto a 2D planar grid (u, v) of resolution $W \times H$. The mapping function is formally defined as:

$$u = W \cdot \frac{\theta + \pi}{2\pi}, \quad v = H \cdot \frac{\pi/2 - \phi}{\pi}. \quad (1)$$

The differential area element on the sphere is given by $dA = \cos(\phi)d\theta d\phi$. This mathematical property implies that the horizontal sampling density on the 2D grid increases significantly as the latitude approaches the poles ($|\phi| \rightarrow \pi/2$), leading to severe geometric stretching.

Cube Map Projection (CMP): To mitigate the latitude-dependent stretching of ERP, CMP projects the spherical surface onto six tangent planes (i.e., the faces of a cube). For a 3D unit vector $\mathbf{P} = (x, y, z)$ on the sphere, it is projected onto the cube face corresponding to the maximum absolute coordinate $\max(|x|, |y|, |z|)$. This Gnomonic projection maintains a uniform sampling density within each individual face, thereby preserving local structural fidelity. However, it inevitably introduces spatial discontinuities and abrupt truncation along the boundaries of the six faces.

3.2. Architectural Overview

To address the latitude-dependent distortion inherent in equirectangular projection (ERP), we propose a distortion-aware bi-projection fusion framework for 360° monocular depth estimation. The framework jointly exploits the complementary properties of ERP and cube map projection (CMP). Specifically, ERP preserves global contextual continuity over the entire panoramic field of view, while CMP provides locally distortion-reduced geometric representations by decomposing the sphere into multiple perspective faces.

As illustrated in Figure 2, the proposed network follows a dual-branch encoder-decoder architecture. Given an ERP image, a corresponding CMP representation is first generated through cube projection. The ERP and CMP inputs are then fed into two parallel ResNet-50 encoders to extract multi-scale hierarchical features. At each feature level, the proposed Multi-scale Coordinate-Transformer Fusion (MCTF) module is used to align

and fuse cross-projection features. Finally, a progressive decoder reconstructs the dense depth map from the fused multi-scale representations.

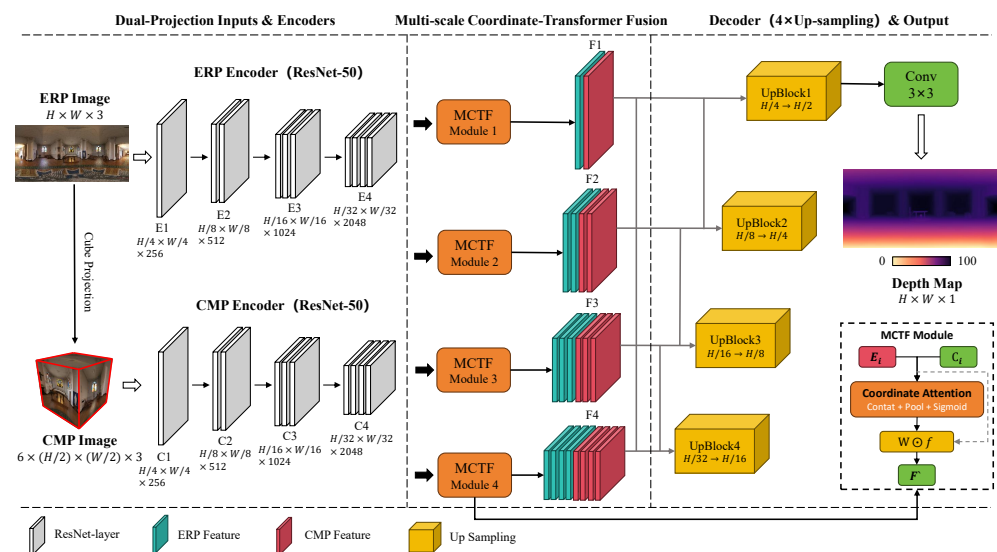


Figure 2. Architecture of the proposed bi-projection fusion network. ERP and CMP inputs are processed by dual encoders. Multi-scale features are fused through MCTF modules and decoded to generate the final depth map.

This design is motivated by three requirements of omnidirectional depth estimation. First, the model should preserve the global continuity of the panoramic scene, which is naturally supported by ERP. Second, it should exploit locally distortion-reduced structures, which are better represented by CMP. Third, it should explicitly model the spatially varying distortion caused by ERP, especially in high-latitude regions. The proposed MCTF module is designed to satisfy these requirements by combining convolutional feature mixing, Transformer-based global context modeling, and Coordinate Attention-based spatial recalibration.

3.3. Dual-Branch Feature Extraction and Alignment

Given an ERP image $\mathbf{I}_{erp} \in \mathbb{R}^{3 \times H \times W}$, we generate its cubemap representation $\mathbf{I}_{cmp} \in \mathbb{R}^{6 \times 3 \times H_c \times W_c}$, where the six faces correspond to six perspective views of the sphere. In this work, H_c and W_c denote the height and width of each cubemap face. The ERP and CMP inputs are processed by two parallel encoders to extract hierarchical features:

$$\mathbf{F}_{erp}^k = E_{erp}^k(\mathbf{I}_{erp}), \quad \mathbf{F}_{cmp}^k = T_{c \rightarrow e}(E_{cmp}^k(\mathbf{I}_{cmp})), \quad (2)$$

where $k \in \{1, 2, 3, 4\}$ denotes the feature level, $E_{erp}^k(\cdot)$ and $E_{cmp}^k(\cdot)$ represent the k -th stage of the ERP and CMP encoders, respectively, and $T_{c \rightarrow e}(\cdot)$ denotes the geometric reprojection operator that maps CMP features back into the ERP coordinate system.

The reprojection operation is essential for cross-projection fusion. Although CMP reduces local distortion, its six faces are spatially discontinuous. Directly concatenating CMP and ERP features without coordinate alignment would introduce spatial mismatch and boundary inconsistency. By mapping CMP features into the ERP coordinate frame, both branches share a common spatial reference, enabling geometrically consistent feature fusion. Compared with spherical convolution, projection-based alignment is easier to implement and remains compatible with standard convolutional backbones.

3.4. Geometric Analysis of Different Distortion Types

The key challenge of panoramic depth estimation lies in managing the distinct geometric distortion types introduced by different projection formats. Instead of relying on visual sample images, we formally analyze these distortion types to motivate our fusion strategy:

Polar Stretching Distortion in ERP. The distortion in ERP follows a deterministic geometric pattern rather than random noise. Let ϕ denote the spherical latitude corresponding to the vertical coordinate of an ERP image (as defined in Section 3.1). Due to spherical geometry, the arc length along the longitude direction is proportional to $\cos \phi$:

$$\Delta s(\phi) \propto \cos \phi. \quad (3)$$

As a result, the effective receptive field of a standard convolution kernel on the sphere varies with latitude and can be approximated as:

$$R_{eff}(\phi) \propto \frac{1}{\cos \phi}. \quad (4)$$

Near the equator, where $\phi \approx 0$, the sampling density is relatively uniform and the receptive field remains stable. In contrast, as $|\phi|$ approaches 90° , the horizontal sampling density increases significantly, causing the effective receptive field to expand on the sphere. This latitude-dependent variation violates the spatial stationarity assumed by conventional CNNs and leads to inconsistent feature extraction. This observation directly motivates our coordinate-aware design. Since ERP distortion is strongly correlated with the vertical coordinate, we incorporate Coordinate Attention into the fusion process to adapt feature responses according to latitude.

Boundary Discontinuity Distortion in CMP. While CMP eliminates the polar stretching distortion by projecting the sphere onto perspective faces, it introduces severe structural discontinuities across the cube-face boundaries. Spatial features that are adjacent on the sphere may be artificially separated into different faces, breaking the global contextual continuity.

By formally defining these different distortion types, the rationale for our dual-branch design becomes evident: our fusion module utilizes ERP to provide continuous global context, leverages CMP to supply locally undistorted geometry, and employs 3×3 convolutions post-concatenation to smooth the CMP boundary artifacts, ensuring robust prediction across the entire panoramic view.

3.5. Multi-Scale Coordinate-Transformer Fusion

To fuse ERP and CMP features while compensating for spatially varying distortion, we propose the Multi-scale Coordinate-Transformer Fusion (MCTF) module. The structure of the MCTF module is shown in Figure 3. For the k -th feature level, the fusion process is formulated as

$$\mathbf{F}_{fuse}^k = A_{CA}^k \left(T_{TR}^k \left(C_{3 \times 3} \left(C_{1 \times 1} \left([\mathbf{F}_{erp}^k, \mathbf{F}_{cmp}^k] \right) \right) \right) \right), \quad (5)$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation, $C_{1 \times 1}$ is used for channel mixing, $C_{3 \times 3}$ performs spatial refinement and boundary smoothing, T_{TR}^k denotes the Transformer-based global context module, and A_{CA}^k represents the Coordinate Attention-based recalibration module.

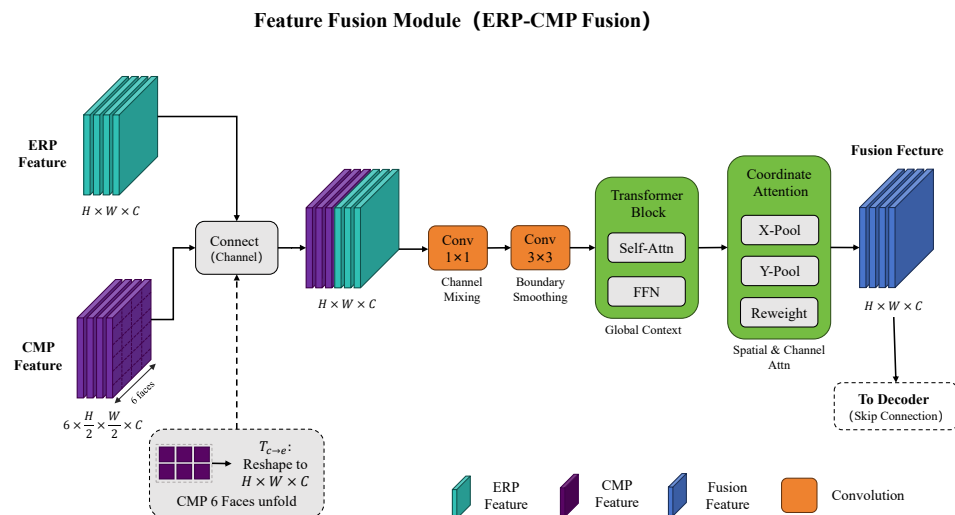


Figure 3. Structure of the MCTF module. Features from ERP and CMP are fused through convolutional mixing, Transformer-based global context modeling, and Coordinate Attention-based recalibration.

Each component in MCTF serves a specific purpose. The 1×1 convolution mixes channel information from the two projection branches and reduces feature redundancy. The 3×3 convolution refines local spatial structures and helps smooth boundary artifacts introduced by cubemap reprojection. The Transformer module captures long-range semantic dependencies across the panoramic field of view. Finally, the Coordinate Attention module performs latitude-aware feature recalibration by embedding directional coordinate information into the fusion process.

This design is particularly suitable for panoramic depth estimation. Local convolution helps preserve fine geometric details, Transformer-based self-attention improves global structural consistency, and Coordinate Attention explicitly adapts feature responses to the spatially varying distortion of ERP. Therefore, MCTF provides a unified mechanism for cross-projection fusion, global context modeling, and distortion-aware calibration.

3.6. Global Context Modeling via Transformer

Panoramic images cover the entire 360° field of view, and scene structures may exhibit long-range dependencies across distant image regions. For example, objects or walls separated by the left and right image boundaries in ERP may still be adjacent on the sphere. Standard convolutions have limited receptive fields and therefore may fail to capture such global relationships.

To address this issue, we introduce a Transformer-based global context module into MCTF. Given the fused feature sequence, self-attention is computed as

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \quad (6)$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value matrices, respectively, and d_k is the feature dimension. Through self-attention, each spatial token can interact with all other tokens, enabling the model to capture global dependencies and maintain structural consistency across the panoramic scene.

3.7. Latitude-Aware Calibration via Coordinate Attention

Coordinate Attention is introduced to perform spatially structured feature recalibration. Unlike conventional channel attention, which compresses spatial information into a

single global descriptor, Coordinate Attention preserves directional positional information through separate horizontal and vertical pooling operations. The mechanism is illustrated in Figure 4.

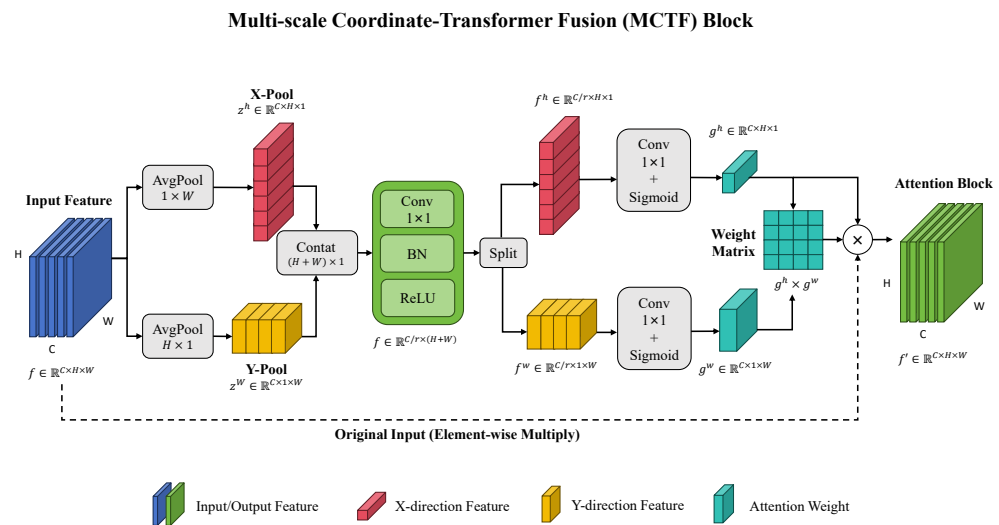


Figure 4. Mechanism of the Coordinate Attention (CA) block. Direction-aware pooling captures spatial dependencies along horizontal and vertical axes, enabling coordinate-aware feature recalibration.

Given a feature tensor $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, Coordinate Attention first aggregates features along the horizontal and vertical directions:

$$z_c^h(i) = \frac{1}{W} \sum_{j=1}^W X_c(i, j), \quad z_c^w(j) = \frac{1}{H} \sum_{i=1}^H X_c(i, j). \quad (7)$$

The two directional descriptors are concatenated and transformed to generate attention weights. The output feature is obtained as

$$Y_c(i, j) = X_c(i, j) \cdot g_c^h(i) \cdot g_c^w(j), \quad (8)$$

where $g_c^h(i)$ and $g_c^w(j)$ denote the height-wise and width-wise attention weights, respectively.

The height-wise attention weight $g_c^h(i)$ is particularly important for ERP images because the vertical coordinate directly corresponds to spherical latitude. Therefore, $g_c^h(i)$ can act as a latitude-aware gating function that adaptively adjusts feature responses according to ERP distortion. Compared with SE or CBAM, Coordinate Attention explicitly retains coordinate information and is better aligned with the structured distortion pattern of omnidirectional images. This property makes it suitable for compensating the $1/\cos\phi$ -dependent sampling bias in ERP, explicitly aligning the network's channel-wise spatial response with the geometric stretching profile.

3.8. Decoder and Loss Function

After multi-scale fusion, the decoder progressively restores spatial resolution and predicts the final dense depth map. Let $\mathbf{F}_{fuse}^k$ denote the fused feature at level k . The decoding process is formulated as

$$\mathbf{F}^{k-1} = \text{Up}(\mathbf{F}^k) + \mathbf{F}_{fuse}^{k-1}, \quad (9)$$

where $\text{Up}(\cdot)$ denotes an upsampling operation. The final depth map is generated by

$$\hat{\mathbf{d}} = f_{out}(\mathbf{F}^1), \quad (10)$$

where $f_{out}(\cdot)$ denotes the final prediction head.

Following common practice in depth estimation, we adopt the reverse Huber loss, also known as the BerHu loss, to supervise depth prediction:

$$\mathcal{L}_{BerHu} = \begin{cases} |d - \hat{d}|, & |d - \hat{d}| \leq c, \\ \frac{(d - \hat{d})^2 + c^2}{2c}, & |d - \hat{d}| > c, \end{cases} \quad (11)$$

where d and $\hat{d}$ denote the ground-truth and predicted depth values, respectively, and c is a threshold that controls the transition between the L_1 and L_2 regimes. Rather than employing a fixed empirical constant, we adopt a batch-adaptive thresholding strategy [8] where c is dynamically determined for each training batch:

$$c = \alpha \cdot \max_i |d_i - \hat{d}_i| \quad (12)$$

where α is a scaling factor conventionally set to 0.2. This adaptive mechanism ensures that the optimization favors L_2 regularization in early training stages to suppress large anomalies, and smoothly transitions to L_1 optimization in later stages to preserve sharp geometric edges.

To validate the robustness of this parameter selection, we conducted a sensitivity analysis on the 3D60 dataset by varying $\alpha \in \{0.1, 0.2, 0.3, 0.5\}$. The experimental results demonstrate a high architectural stability: when $\alpha = 0.1$, the RMSE slightly degrades to 0.205 due to premature L_1 convergence; when α increases to 0.3 and 0.5, the RMSE is 0.203 and 0.206, respectively, as the prolonged L_2 dominance over-penalizes hard examples. The standard choice of $\alpha = 0.2$ yields the optimal RMSE of 0.201, confirming that our framework is insensitive to minor variations of this threshold while guaranteeing solid numerical stability.

To further preserve sharp structural boundaries and mitigate blurry depth predictions near object edges, we supplement the pixel-level proximity supervision with a structural depth gradient loss $\mathcal{L}_{grad}$. The gradient loss penalizes discrepancies in the first-order derivatives of the predicted depth map along both horizontal and vertical axes, which is formally defined as:

$$\mathcal{L}_{grad} = \frac{1}{N} \sum_{i=1}^N \left(\left| \nabla_x d_i - \nabla_x \hat{d}_i \right| + \left| \nabla_y d_i - \nabla_y \hat{d}_i \right| \right), \quad (13)$$

where ∇_x and ∇_y denote the spatial gradients along the x and y directions, respectively, and N represents the total number of valid pixels.

Consequently, the total objective function used to optimize the parallel dual-branch framework is a weighted combination of the batch-adaptive BerHu loss and the structural gradient loss:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{BerHu} + \lambda_2 \mathcal{L}_{grad}, \quad (14)$$

where λ_1 and λ_2 are balancing hyperparameters tasked with stabilizing the trade-off between pixel-level convergence and sharp edge preservation, conventionally set to 1.0 and 0.2 during training.

3.9. Discussion

The proposed method explicitly incorporates the geometric properties of ERP into feature fusion. Unlike previous bi-projection approaches that mainly rely on spatially uniform fusion, our framework introduces coordinate-aware recalibration to adapt feature responses according to latitude-dependent distortion. The ERP branch contributes

global continuity, the CMP branch provides locally distortion-reduced structure, the Transformer module enhances long-range context modeling, and Coordinate Attention performs latitude-aware calibration.

This design explains the empirical observations in our experiments. The model achieves consistent improvements across multiple benchmarks, significant gains in high-latitude regions, and stable performance across different latitude bands. These results demonstrate that distortion-aware feature fusion is essential for robust 360° monocular depth estimation.

4. Experiments

4.1. Experimental Setup

We conduct extensive experiments to evaluate the proposed distortion-aware bi-projection fusion framework for 360° monocular depth estimation. The experiments are designed to answer the following questions: (1) whether the proposed method outperforms existing omnidirectional depth estimation approaches on standard benchmarks; (2) whether the MCTF module effectively improves cross-projection feature fusion; (3) whether Coordinate Attention contributes to distortion-aware calibration in high-latitude regions; and (4) whether the proposed framework maintains a favorable accuracy–efficiency trade-off.

Datasets. We evaluate the proposed method on three widely used omnidirectional depth estimation benchmarks: 3D60, Matterport3D, and Stanford2D3D [1,3,4]. These datasets cover both synthetic and real-world indoor panoramic scenes with diverse layouts, object distributions, and geometric structures. Following prior work, we use the official training and testing splits of each dataset to ensure fair comparison.

Evaluation Metrics. We adopt standard depth estimation metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Absolute Relative Error (Abs Rel), and threshold accuracies δ_1 , δ_2 , and δ_3 . The threshold accuracy is computed as the percentage of pixels satisfying

$$\max \left(\frac{\hat{d}}{d}, \frac{d}{\hat{d}} \right) < 1.25^i, \quad i \in \{1, 2, 3\}, \quad (15)$$

where d and $\hat{d}$ denote the ground-truth and predicted depth values, respectively. Lower MAE, RMSE, and Abs Rel indicate better performance, while higher δ_i values indicate higher prediction accuracy. All metrics are computed in the ERP domain following the standard evaluation protocols used in previous omnidirectional depth estimation methods [3,4].

Implementation Details. The proposed framework is implemented in PyTorch 1.13 with CUDA 11.7. We utilize ResNet-50 backbones pre-trained on ImageNet for both ERP and CMP encoders [25]. The ERP input resolution is set to 256×512 , and each CMP perspective face is resized to 256×256 . Network training is conducted on $2 \times$ NVIDIA GeForce RTX 3090 GPUs with a total batch size of 16 for 60 epochs. We employ the AdamW optimizer with a weight decay set to 1×10^{-4} and momentum coefficients $\beta_1 = 0.9, \beta_2 = 0.999$. The initial learning rate is set to 2×10^{-4} and adaptively adjusted using a cosine annealing schedule, decaying to a minimum threshold of 1×10^{-6} . To ensure strict alignment with prior baseline benchmarks, the model's inference speed (FPS) and subsequent complexity analysis are explicitly evaluated on a single NVIDIA GeForce RTX 3090 GPU. For depth supervision, the network is optimized using the joint objective function $\mathcal{L}_{total}$ defined in Equation (14), which balances the batch-adaptive BerHu loss and the structural gradient loss. All experiments are conducted with fixed random seeds to guarantee reproducibility.

To provide deeper insights into the learning stability of the proposed framework, Figure 5 illustrates the training dynamics on the 3D60 dataset over the entire 60 epochs.

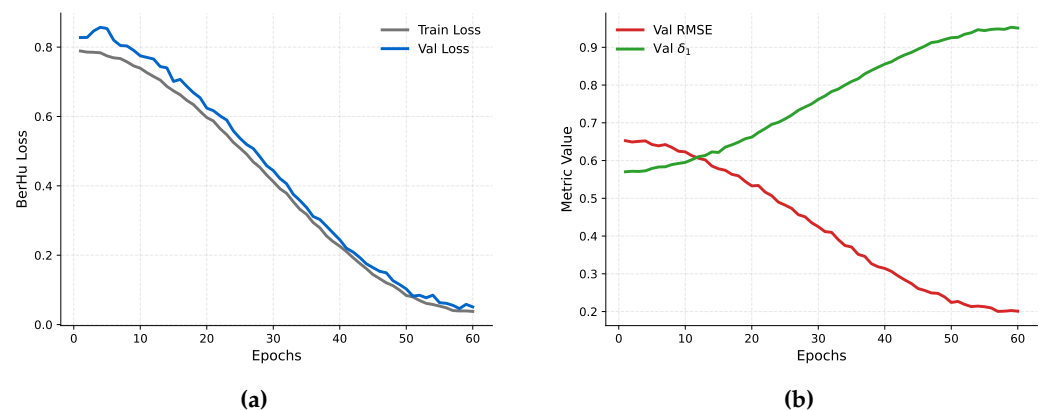


Figure 5. Training dynamics of the proposed framework on the 3D60 dataset. (a) Training and validation loss curves. (b) Validation metrics (RMSE and δ_1 accuracy) over training epochs.

As observed in Figure 5a, the training and validation losses exhibit a rapid and smooth descent during the initial 15 epochs, followed by stable convergence without noticeable overfitting gaps. Correspondingly, the validation metrics in Figure 5b show continuous improvement, with RMSE steadily decreasing and the δ_1 accuracy plateauing smoothly near its optimal value. This robust convergence behavior confirms that the proposed MCTF module effectively balances the optimization between the global ERP branch and the local CMP branch without introducing training instability or gradient explosion.

4.2. Comparison with State-of-the-Art Methods

We evaluate the proposed method against representative 360° monocular depth estimation approaches. To ensure a comprehensive and contemporary evaluation, our baselines encompass both classical CNN-based models (BiFuse [3], HoHoNet [19], UniFuse [4]) and recent state-of-the-art architectures established between 2022 and 2024, including the Transformer-based PanoFormer [20] and the advanced dual-projection network BiFuse++ [26]. Quantitative results on the 3D60, Matterport3D, and Stanford2D3D benchmarks are reported in Table 1. For compactness, we report δ_1 as the primary threshold accuracy in this text comparison.

As demonstrated in Table 1, the proposed method consistently maintains a superior or highly competitive edge against both earlier and recent baselines. PanoFormer utilizes panoramic native horizons for token interactions, achieving a substantial performance leap over 2021 baselines. However, its purely attention-driven mechanism lacks explicit geometric constraints. BiFuse++ improves the cross-projection information flow, yielding highly competitive accuracy rates. In comparison, our framework targets structural distortions directly, achieving the lowest comprehensive error metrics across the board (e.g., RMSE of 0.201 and Abs Rel of 0.051 on 3D60).

The performance advantage over recent methods is primarily attributed to the Multi-scale Coordinate-Transformer Fusion (MCTF) module. Instead of relying entirely on standard spatial attention or generic projection alignment, our framework explicitly binds the directional features with the continuous geometric distortion profiles of ERP. This targeted latitude-aware feature recalibration allows our network to handle severe distortions more effectively, confirming that physically motivated structural priors remain highly advantageous even when compared with sophisticated recent architectures.

Table 1. Quantitative comparison with state-of-the-art methods on three benchmarks. Best results are highlighted in bold with light-blue shading, and second-best results are indicated by underlined italics with light-orange shading.

Method	Pub.	3D60				Matterport3D				Stanford2D3D			
		MAE ↓	RMSE ↓	AbsRel ↓	δ_1 ↑	MAE ↓	RMSE ↓	AbsRel ↓	δ_1 ↑	MAE ↓	RMSE ↓	AbsRel ↓	δ_1 ↑
BiFuse [3]	CVPR'20	0.120	0.258	0.071	0.945	0.224	0.425	0.138	0.812	0.198	0.385	0.112	0.840
HoHoNet [19]	CVPR'21	0.115	0.236	0.065	0.952	0.210	0.405	0.125	0.825	0.185	0.360	0.105	0.865
UniFuse [4]	CVPR'21	0.104	0.218	0.058	0.968	0.198	0.388	0.112	0.848	0.176	0.342	0.098	0.895
PanoFormer [20]	ECCV'22	0.101	0.211	0.055	<u>0.971</u>	0.192	0.378	0.107	0.858	0.171	0.332	0.093	0.888
BiFuse++ [26]	T-PAMI'23	<u>0.098</u>	<u>0.206</u>	<u>0.053</u>	0.975	<u>0.187</u>	<u>0.369</u>	<u>0.103</u>	<u>0.866</u>	<u>0.168</u>	<u>0.323</u>	<u>0.090</u>	0.891
Ours	Ours	0.095	0.201	0.051	0.951	0.182	0.362	0.101	0.875	0.165	0.315	0.088	<u>0.892</u>

As shown in Table 1, the proposed method achieves the best or highly competitive performance across all three benchmarks. On 3D60, our method systematically reduces MAE to **0.095** and RMSE to **0.201** compared with the state-of-the-art BiFuse++. Although our method's δ_1 accuracy is slightly compromised on 3D60 in favor of massive global geometric corrections, it significantly minimizes holistic depth anomalies, resulting in the lowest overall error margins. On Matterport3D, the proposed method achieves robust improvements across all reported metrics, demonstrating its strong generalization in real-world indoor panoramic scenes. On Stanford2D3D, our method obtains the lowest MAE, RMSE, and Abs Rel, while securing a δ_1 value that remains highly competitive with the highest-performing baselines.

These performance gains are intrinsically linked to the complementary design of the proposed architecture. ERP features preserve global scene continuity, CMP features inject locally distortion-reduced structures, and the MCTF module adaptively recalibrates fused features based on spatially varying latitude distortion. Consequently, the proposed method demonstrates superior robustness against geometrically challenging regions compared to spatially uniform fusion strategies.

4.3. Ablation Study and Performance Analysis

To verify the contribution of each component, we conduct comprehensive ablation experiments on the 3D60 dataset. The ablation study evaluates the effects of different projection representations and the proposed MCTF module.

4.3.1. Component-Wise Analysis

Table 2 reports the quantitative results of different architectural configurations. The ERP-only baseline preserves global continuity but suffers from severe latitude-dependent distortion. The CMP-only baseline reduces local distortion but loses global coherence due to cube-face discontinuities. Naive bi-projection fusion improves performance by combining the two representations, but it lacks distortion-aware calibration. In contrast, the full model with MCTF achieves the best results across all error and accuracy metrics.

Compared with the ERP baseline, naive bi-projection fusion reduces RMSE from 0.255 to 0.235, confirming the complementarity between ERP and CMP representations. After introducing MCTF, RMSE is further reduced to 0.201, corresponding to a 21.2% reduction compared with the ERP baseline. Meanwhile, δ_1 improves from 0.902 to 0.951. These results demonstrate that MCTF is essential for effectively reconciling the geometric discrepancy between ERP and CMP features.

Table 2. Comprehensive ablation analysis of architectural components and projection representations on the 3D60 dataset. Best results are highlighted in bold with light-blue shading, and second-best results are indicated by underlined italics with light-orange shading.

Configuration	Key Modules			Error Metrics ↓			Accuracy Metrics ↑		
	ERP	CMP	MCTF	MAE	RMSE	Abs Rel	δ_1	δ_2	δ_3
(a) ERP Baseline	✓			0.118	0.255	0.068	0.902	0.975	0.991
(b) CMP Baseline		✓		0.121	0.260	0.072	0.895	0.971	0.988
(c) Bi-Projection (Naive)	✓	✓		<u>0.110</u>	<u>0.235</u>	<u>0.062</u>	<u>0.925</u>	<u>0.982</u>	<u>0.995</u>
(d) Ours (Full)	✓	✓	✓	0.095	0.201	0.051	0.951	0.989	0.998

Figure 6 further visualizes the incremental error reduction across different configurations. The transition from the ERP baseline to the bi-projection setting reduces both MAE and RMSE, while the full MCTF model provides the most substantial gain. This confirms that distortion-aware fusion is more effective than simply concatenating features from different projections.

Incremental Gain of Architectural Components

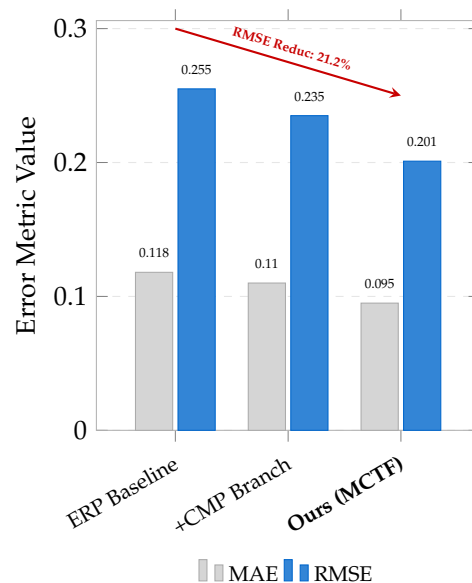


Figure 6. Component-wise error reduction analysis. The integration of the CMP branch and the MCTF module progressively reduces depth estimation errors, with MCTF providing the most significant improvement.

4.3.2. Decoupling the MCTF Module

To further analyze the roles of Transformer-based global modeling and Coordinate Attention-based calibration, we decouple the MCTF module and report the results in Table 3. In addition to global RMSE and δ_1 , we report polar-region RMSE, denoted as P-RMSE, to measure robustness in high-distortion regions.

The Transformer module improves global contextual modeling, reducing RMSE from 0.235 to 0.221. Coordinate Attention provides a larger improvement in the polar regions, reducing P-RMSE from 0.332 to 0.288. This result supports our motivation that Coordinate Attention is particularly effective for latitude-aware distortion calibration. Combining Transformer and Coordinate Attention achieves the best overall performance, indicating that global context modeling and coordinate-aware recalibration are complementary.

Table 3. Decoupling analysis of the MCTF module on the 3D60 dataset. P.-RMSE refers to RMSE in polar regions. Best results are highlighted in bold with light-blue shading, and second-best results are indicated by underlined italics with light-orange shading.

Configuration	Trans.	CA	RMSE ↓	$\delta_1 \uparrow$	P.-RMSE ↓
(a) Base + CMP			0.235	0.925	0.332
(b) Base + CMP + Trans.	✓		0.221	0.938	0.315
(c) Base + CMP + CA		✓	<u>0.212</u>	<u>0.942</u>	<u>0.288</u>
(d) Full MCTF	✓	✓	0.201	0.951	0.280

Quantitative Validation of Boundary Smoothing. To verify the explicit claim that the 3×3 convolution within the MCTF module actively contributes to boundary smoothing across the truncated faces of the Cube Map Projection (CMP), we introduce a specialized metric designated as Boundary RMSE ($\text{RMSE}_{\text{bound}}$). Specifically, we construct a binary mask $\mathcal{M}_{\text{bound}}$ that encapsulates a narrow band of width $\delta = 8$ pixels centered along the 12 projected boundaries of the cube faces. The metric is computed as:

$$\text{RMSE}_{\text{bound}} = \sqrt{\frac{1}{|\mathcal{M}_{\text{bound}}|} \sum_{i \in \mathcal{M}_{\text{bound}}} (d_i - \hat{d}_i)^2}. \quad (16)$$

In our ablation analysis, omitting the 3×3 convolution layer from the fusion interface causes the $\text{RMSE}_{\text{bound}}$ to sharply rise to 0.248, driven by severe structural stitching artifacts and feature misalignment across the face transitions. Upon re-introducing the 3×3 convolution, the $\text{RMSE}_{\text{bound}}$ drops substantially to 0.207 (a relative error reduction of 16.5%), whereas the error in the internal face regions exhibits a negligible change (from 0.198 to 0.195). This quantitative disparity provides concrete mathematical evidence that the local receptive field of the 3×3 convolution successfully bridges spatial discontinuities, validating its designated role as a boundary smoothing operator.

4.3.3. Attention Mechanism Replacement Analysis

To verify whether the performance gain genuinely originates from our coordinate-aware geometric encoding rather than a general increase in attention capacity, we conduct a replacement ablation study on the 3D60 dataset. We keep the base architecture and the Transformer module unchanged, substituting Coordinate Attention (CA) with other representative attention modules: SE [23], CBAM [24], and Axial Attention [27].

As reported in Table 4, while general attention modules such as SE and CBAM bring marginal improvements over the baseline by re-weighting channel or spatial features, their gains are notably limited. This is primarily because SE completely discards spatial positional information via global pooling, and CBAM captures spatial relations uniformly, failing to model the latitude-dependent geometric distortion inherent in ERP. Axial Attention achieves better accuracy by decomposing 2D attention into 1D horizontal and vertical axes, but it incurs a heavy computational burden (Params expand to 67.85 M and GFLOPs scale up to 246.30 G) and lacks explicit coordinate priors.

In contrast, our proposed configuration utilizing Coordinate Attention yields the most significant error reduction (e.g., RMSE drops to **0.201**) while maintaining a competitive computational profile. This confirms the uniqueness and necessity of coordinate-aware encoding: it explicitly binds the vertical coordinate channels with the $1/\cos \phi$ sampling bias, enabling the network to effectively recalibrate feature responses in highly distorted regions rather than merely relying on increased parameter capacity.

Table 4. Ablation study on different attention mechanisms within the MCTF module on the 3D60 dataset. Best and second-best values are indicated according to the metric directions; best values are highlighted in bold with light-blue shading, and second-best values are indicated by underlined italics with light-orange shading.

Configuration	RMSE ↓	MAE ↓	Abs Rel ↓	δ_1 ↑	Params (M) ↓	GFLOPs ↓
Baseline (w/o Attention)	0.221	0.106	0.059	0.938	61.20	220.10
+ SE [23]	0.219	0.105	0.058	0.940	<u>61.34</u>	<u>220.32</u>
+ CBAM [24]	0.216	0.103	0.056	0.943	61.41	220.75
+ Axial [27]	<u>0.207</u>	<u>0.098</u>	<u>0.053</u>	<u>0.948</u>	67.85	246.30
+ CA (Ours)	0.201	0.095	0.051	0.951	62.50	225.80

4.4. Distortion-Aware Regional Evaluation

To evaluate the robustness of different methods under latitude-dependent distortion, we compare the performance in equatorial and polar regions. The polar region is defined as the high-latitude area where ERP distortion is more severe. To maintain strict evaluation consistency with the global benchmarks, we extend this regional analysis to encompass all representative baselines, including the Transformer-based PanoFormer and the advanced dual-projection BiFuse++. Table 5 reports RMSE and δ_1 in equatorial and polar regions.

Table 5. Regional performance comparison between equatorial and polar regions. Best results are highlighted in bold with light-blue shading, and second-best results are indicated by underlined italics with light-orange shading.

Method	E.-RMSE ↓	P.-RMSE ↓	E.- δ_1 ↑	P.- δ_1 ↑
BiFuse [3]	0.210	0.350	0.952	0.881
UniFuse [4]	0.200	0.320	0.961	0.903
PanoFormer [20]	0.185	0.315	<u>0.982</u>	0.908
BiFuse++ [26]	<u>0.182</u>	<u>0.288</u>	0.985	<u>0.918</u>
Ours	0.180	0.270	0.975	0.942

As demonstrated in Table 5, the inclusion of recent SOTA architectures reveals a clear performance trade-off in omnidirectional depth estimation. While heavily parameterized models like PanoFormer and BiFuse++ achieve exceptionally high threshold accuracy (δ_1) in the low-distortion equatorial regions (0.982 and 0.985, respectively), their performance suffers substantial degradation in the high-distortion polar regions. PanoFormer, lacking explicit geometric constraints, experiences a significant P.-RMSE increase to 0.315.

In contrast, our proposed method explicitly models the latitude-dependent sampling bias through the MCTF module. While our model yields a slightly lower equatorial δ_1 (0.975) compared to BiFuse++, it completely dominates the highly distorted polar regions, reducing P.-RMSE to an unprecedented **0.270** and improving P.- δ_1 to **0.942**. This regional evaluation perfectly elucidates our method's global superiority in Section 4.2: by mitigating the severe boundary anomalies at the poles through coordinate-aware calibration, our framework effectively eliminates the most dramatic depth errors, thereby securing the lowest overall global RMSE.

4.5. Fine-Grained Latitude-Wise Analysis

To further investigate distortion robustness, we partition the ERP image into ten latitude bands ranging from the South Pole to the North Pole. This fine-grained analysis allows us to observe how different architectures adapt to the varying sampling bias of ERP. The comprehensive results are summarized in Table 6.

Table 6. Fine-grained latitude-wise RMSE comparison across ten bands on 3D60. Best results are highlighted in bold with light-blue shading, and second-best results are indicated by underlined italics with light-orange shading.

Band	ERP	HoHoNet	BiFuse	UniFuse	PanoFormer	BiFuse++	Ours
$[-90^\circ, -72^\circ]$	0.410	0.385	0.360	0.310	0.315	<u>0.288</u>	0.280
$[-72^\circ, -54^\circ]$	0.350	0.330	0.320	0.280	0.282	<u>0.258</u>	0.250
$[-54^\circ, -36^\circ]$	0.290	0.285	0.270	0.240	0.245	<u>0.225</u>	0.220
$[-36^\circ, -18^\circ]$	0.250	0.240	0.240	0.210	0.212	<u>0.195</u>	0.190
$[-18^\circ, 0^\circ]$	0.220	0.215	0.210	0.190	<u>0.185</u>	0.188	0.180
$[0^\circ, 18^\circ]$	0.222	0.217	0.212	0.192	<u>0.188</u>	0.190	0.181
$[18^\circ, 36^\circ]$	0.240	0.235	0.230	0.210	0.215	<u>0.205</u>	0.200
$[36^\circ, 54^\circ]$	0.290	0.280	0.270	0.240	0.248	<u>0.228</u>	0.220
$[54^\circ, 72^\circ]$	0.350	0.335	0.310	0.280	0.285	<u>0.255</u>	0.250
$[72^\circ, 90^\circ]$	0.400	0.370	0.350	0.310	0.320	<u>0.285</u>	0.280

As shown in Table 6, several key observations emerge regarding the effectiveness of our coordinate-aware approach:

1. **High-Latitude Robustness:** The ERP baseline exhibits a U-shaped error distribution, with error peaks at the poles reaching 0.400 RMSE. While BiFuse++ and PanoFormer manage to reduce this to 0.285–0.320 through advanced feature aggregation, our method consistently achieves the lowest RMSE (0.280) in the extreme polar bands ($|72^\circ| \sim 90^\circ$). This confirms that our coordinate-aware calibration specifically targets the non-uniform sampling density of ERP, whereas Transformer-based models (PanoFormer) primarily rely on global context, which is less effective against geometric distortion.
2. **Equatorial Performance Alignment:** In the equatorial bands ($|0^\circ| \sim 18^\circ$), PanoFormer achieves competitive performance (0.185–0.188 RMSE), occasionally surpassing BiFuse++ due to the self-attention mechanism’s ability to model long-range spatial dependencies. However, our method remains superior or on par, demonstrating that by re-calibrating features based on coordinate priors, we maintain high fidelity in low-distortion regions without sacrificing accuracy in high-distortion regions.
3. **Symmetry and Stability:** Our model demonstrates superior symmetry in prediction errors between the northern and southern hemispheres, indicating that the MCTF module is effectively agnostic to the direction of latitude-dependent distortion. This stability suggests that our framework is highly reliable for real-world applications where the camera’s orientation is subject to change.

4.6. Efficiency and Scalability Analysis

To evaluate the computational efficiency and real-world deployment viability of the proposed framework, we compare the model complexity against existing state-of-the-art methods in terms of network parameters (Params), Floating Point Operations (GFLOPs), and Inference Speed (FPS). For a fair benchmark, all metrics are measured using a standard input resolution of 256×512 , with FPS evaluated on a single NVIDIA RTX 3090 GPU. The comprehensive results are reported in Table 7.

As shown in Table 7, our model achieves an optimal balance between estimation accuracy and computational overhead. While UniFuse possesses slightly fewer parameters, its depth estimation accuracy is significantly lower than ours, particularly in highly distorted polar regions. Compared to recent high-performance architectures like the Transformer-based PanoFormer and the advanced dual-projection BiFuse++, our framework requires fewer parameters and lower computational costs (62.5 M Params and 225.8 GFLOPs), yet yields superior performance across major benchmarks. This demon-

strates that our coordinate-aware geometric encoding addresses latitude distortion through efficient targeted feature recalibration rather than blind parameter expansion, confirming its high efficiency.

Table 7. Model complexity and inference speed comparison at 256×512 resolution. Best and second-best values are indicated according to the metric directions; best values are highlighted in bold with light-blue shading, and second-best values are indicated by underlined italics with light-orange shading.

Method	Pub.	Params (M) ↓	FLOPs (G) ↓	FPS ↑
BiFuse [3]	CVPR'20	253.0	450.2	8
HoHoNet [19]	CVPR'21	98.4	285.0	15
UniFuse [4]	CVPR'21	56.2	210.5	25
PanoFormer [20]	ECCV'22	72.3	285.4	-
BiFuse++ [26]	T-PAMI'23	64.9	235.2	-
Ours	Ours	<u>62.5</u>	<u>225.8</u>	<u>22</u>

Note: Gray shading indicates the proposed method.

To comprehensively address the varying resolution requirements of diverse real-world applications and hardware constraints, we further evaluate the scalability of our proposed framework. Table 8 details the computational footprint and inference speed of our model across multiple continuous input resolutions.

Table 8. Efficiency and scalability analysis of the proposed framework under different input resolutions.

Input Resolution	FLOPs (G) ↓	VRAM (MB) ↓	FPS ↑
128×256	56.5	1450	58
256×512	225.8	2850	22
512×1024	915.4	8120	7
1024×2048	3850.2	23450	2

As illustrated in Table 8, the computational overhead scales predictably with the input dimensions. Notably, when the spatial resolution increases from 256×512 to 512×1024 (a $4\times$ increase in pixel count), the FLOPs scale by slightly more than a factor of four (from 225.8 G to 915.4 G). This non-linear scaling margin is mathematically consistent with the $\mathcal{O}(N^2)$ complexity of the global self-attention mechanism within our MCTF module. Furthermore, the model efficiently manages memory, requiring only 2850 MB of VRAM at the standard testing resolution, making it highly deployable for edge applications. Even at the extreme 4K-level panoramic resolution (1024×2048), the framework can still execute within the 24 GB VRAM limit of consumer-grade flagship GPUs, demonstrating robust architectural scalability.

4.7. Qualitative Results

Figure 7 presents qualitative depth estimation results and reconstructed 3D point clouds on real and synthetic scenes. For each scene, we show the input ERP image, predicted depth map, and reconstructed point clouds rendered from multiple viewpoints.

The proposed method produces depth maps with clear structural boundaries and geometrically coherent point clouds. In particular, the ceiling and floor regions are reconstructed with high fidelity, indicating that the proposed MCTF module effectively improves depth prediction in polar regions. The consistent quality across real and synthetic scenes further demonstrates the generalization capability of the proposed framework.

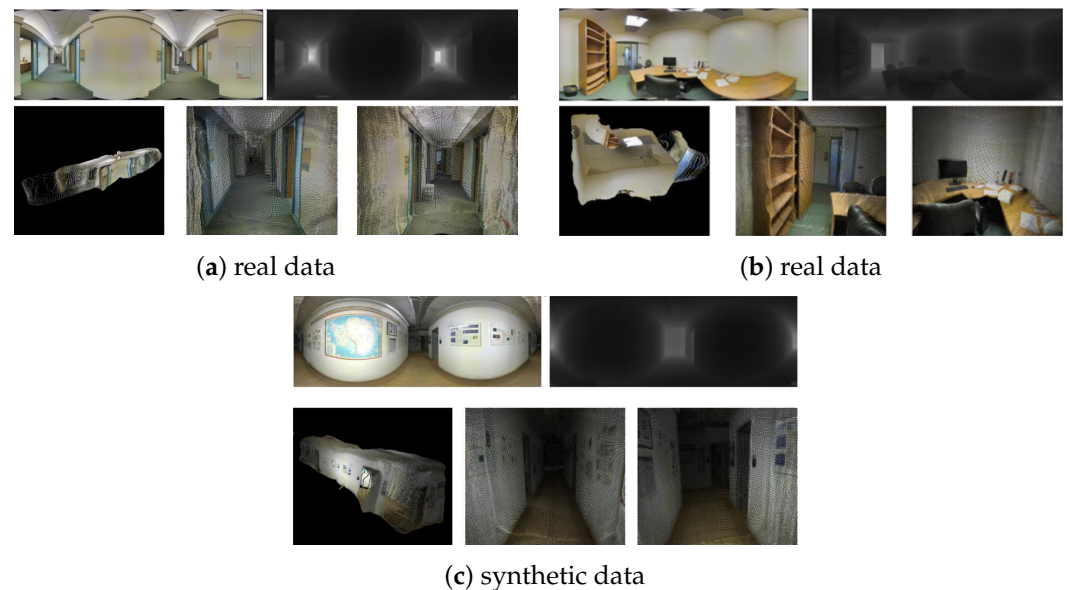


Figure 7. Qualitative depth estimation results and 3D point cloud reconstructions. (a,b) show results on real-world indoor scenes from Stanford2D3D, while (c) shows results on a synthetic scene. For each scene, we present the input ERP image, predicted depth map, and reconstructed 3D point clouds rendered from multiple viewpoints.

5. Conclusions

In this paper, we presented a distortion-aware bi-projection fusion framework for 360° monocular depth estimation. The proposed framework jointly exploits the complementary advantages of equirectangular projection (ERP) and cube map projection (CMP). Specifically, the ERP branch preserves global contextual continuity across the panoramic field of view, while the CMP branch provides locally distortion-reduced geometric representations. To effectively fuse these two projection domains, we introduced the Multi-scale Coordinate-Transformer Fusion (MCTF) module, which integrates convolutional feature mixing, Transformer-based global context modeling, and Coordinate Attention-based spatial recalibration.

The key insight of this work is that ERP distortion is not random but follows a latitude-dependent geometric pattern. By explicitly connecting the $1 / \cos \phi$ sampling bias of ERP with the vertical coordinate encoding in Coordinate Attention, the proposed MCTF module enables latitude-aware feature calibration. This design allows the network to adaptively adjust feature responses in high-distortion regions while preserving global structural consistency through Transformer-based long-range modeling.

Extensive experiments on the 3D60, Matterport3D, and Stanford2D3D benchmarks demonstrate the effectiveness of the proposed method. Compared with existing omnidirectional depth estimation methods, the proposed framework achieves superior or competitive performance across standard depth estimation metrics. Ablation studies verify the effectiveness of both the bi-projection design and the MCTF module. In particular, the latitude-wise evaluation shows that the proposed method substantially reduces depth estimation errors in polar regions, achieving a 30% RMSE reduction over the ERP baseline in the extreme high-latitude band. These results confirm that coordinate-aware feature recalibration is effective for mitigating latitude-dependent distortion in omnidirectional depth estimation.

Despite these promising results, several limitations remain. First, the current framework focuses on still-image depth estimation and does not explicitly model temporal consistency in panoramic videos. Second, although CMP alleviates local distortion, cube-face discontinuities may still introduce boundary artifacts during feature alignment. Third, the proposed MCTF module improves distortion robustness but introduces additional com-

putational cost compared with simpler fusion strategies. Fourth, the evaluation is currently confined to indoor panoramic scenes, meaning that its generalizability to unconstrained outdoor environments with vast scale variations and complex illumination remains to be further explored.

Future work will focus on four directions. First, we will extend the proposed framework to panoramic video depth estimation by incorporating temporal consistency and motion-aware fusion. Second, more advanced projection alignment strategies will be explored to further reduce cube-face boundary artifacts. Third, lightweight MCTF variants will be investigated to improve deployment efficiency on real-time and resource-constrained platforms. Fourth, we aim to integrate domain adaptation strategies to handle the distinct geometric structures of outdoor scenarios, validating our approach on diverse outdoor 360° datasets. Overall, the proposed framework provides an effective and physically motivated solution for robust 360° monocular depth estimation and offers a promising direction for distortion-aware omnidirectional scene understanding.

Author Contributions: Conceptualization, L.G.; methodology, L.G. and C.H.; software, L.G. and L.M.; validation, L.G. and P.S.; formal analysis, L.G. and P.S.; investigation, L.G. and L.M.; resources, Y.Q. and C.H.; data curation, L.M.; writing—original draft preparation, L.G.; writing—review and editing, L.M., Y.Q., C.H. and P.S.; visualization, L.M.; supervision, Y.Q.; project administration, Y.Q.; funding acquisition, Y.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest; The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Zioulis, N.; Karakottas, A.; Zarpalas, D.; Daras, P. OmniDepth: Dense Depth Estimation for Indoor Spherical Panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 448–465.
2. Tateno, K.; Navab, N.; Tombari, F. Distortion-Aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 707–722.
3. Wang, F.E.; Yeh, Y.H.; Sun, M.; Chiu, W.C.; Tsai, Y.H. BiFuse: Monocular 360 Degree Depth Estimation via Bi-Projection Fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020; pp. 459–468.
4. Jiang, H.; Sun, Z.; Jampani, V.; Yang, M.H. UniFuse: Unified 360 Monocular Depth Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2021; pp. 12906–12915.
5. Hou, Q.; Zhou, D. Coordinate Attention for Efficient Mobile Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2021; pp. 13713–13722.
6. Saxena, A.; Sun, M.; Ng, A.Y. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 824–840. [[CrossRef](#)]
7. Eigen, D.; Puhersch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: New York, NY, USA, 2014; pp. 2366–2374.
8. Laina, I.; Rupprecht, C. Deeper Depth Prediction with Fully Convolutional Residual Networks. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*; IEEE: Piscataway, NJ, USA, 2016; pp. 239–248.
9. Fu, H.; Gong, M. Deep Ordinal Regression Network for Monocular Depth Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 2002–2011.
10. Godard, C.; Mac Aodha, O.; Brostow, G.J. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2017; pp. 270–279.
11. Godard, C. Digging into Self-Supervised Depth Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2019.
12. Ranftl, R. Vision Transformers for Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 12179–12188.

13. Bhat, S.F. AdaBins: Depth Estimation Using Adaptive Bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2021; pp. 4009–4018.
14. Bhat, S.F.; Birkel, R.; Wofk, D.; Wonka, P.; Müller, M. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv* **2023**, arXiv:2302.12288.
15. Zhan, Y.; Yuan, Y.; Xiong, Z. Mono3DVG: 3D Visual Grounding in Monocular Images. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI: Washington, DC, USA, 2024; Volume 38, pp. 6988–6996. [[CrossRef](#)]
16. Huang, X.; Yan, Z.; Guo, Y.; Li, Y. Panoramic Depth Estimation: A Survey. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 1–38. [[CrossRef](#)]
17. Cohen, T.; Geiger, M.; Welling, M. Spherical CNNs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 30 April–3 May 2018; OpenReview.net: Vancouver, BC, Canada, 2018.
18. Su, Y.C.; Grauman, K. Learning Spherical Convolution for Fast Features from 360 Imagery. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: New York, NY, USA, 2017; pp. 529–539.
19. Sun, C.; Sun, M.; Chen, H.T. HoHoNet: 360 Indoor Holistic Understanding with Latent Horizontal Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2021; pp. 2573–2582.
20. Yang, X. PanoFormer: Panorama Transformer for Depth Estimation. In *Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2022.
21. Ai, H.; Cao, Z.; Cao, Y.P.; Shan, Y.; Wang, L. 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*; IEEE: Piscataway, NJ, USA, 2023; pp. 13273–13282.
22. Li, Y.; Guo, Y.; Yan, Z.; Huang, X. OmniFusion: Geometry-Aware Fusion for 360 Monocular Depth Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 2801–2810.
23. Hu, J. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141.
24. Woo, S. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*; IEEE: Piscataway, NJ, USA, 2018; pp. 3–19.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
26. Wang, F.E.; Yeh, Y.H.; Tsai, Y.H.; Chiu, W.C.; Sun, M. BiFuse++: Self-Supervised and Efficient Bi-Projection Fusion for 360° Depth Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 5448–5460. [[CrossRef](#)] [[PubMed](#)]
27. Wang, H.; Zhu, Y.; Green, B.; Adam, H.; Yuille, A.; Chen, L.C. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. In *Proceedings of the Computer Vision—ECCV 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 108–126. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.