MCPDepth: Practical Omnidirectional Depth Estimation from Multiple Cylindrical Panoramas via Stereo Matching

Feng Qiao[1], Zhexiao Xiong[1], Xinge Zhu[2], Yuexin Ma[3], Qiumeng He[4], Nathan Jacobs[1]
[1]Washington University in St. Louis [2]The Chinese University of Hong Kong
[3]ShanghaiTech University [4]University of California, Los Angeles

Abstract

Omnidirectional depth estimation presents a significant challenge due to the inherent distortions in panoramic images. Despite notable advancements, the impact of projection methods remains underexplored. We introduce Multi-Cylindrical Panoramic Depth Estimation (MCPDepth), a novel two-stage framework designed to enhance omnidirectional depth estimation through stereo matching across multiple cylindrical panoramas. MCPDepth initially performs stereo matching using cylindrical panoramas, followed by a robust fusion of the resulting depth maps from different views. Unlike existing methods that rely on customized kernels to address distortions, MCPDepth utilizes standard network components, facilitating seamless deployment on embedded devices while delivering exceptional performance. To effectively address vertical distortions in cylindrical panoramas, MCPDepth incorporates a circular attention module, significantly expanding the receptive field beyond traditional convolutions. We provide a comprehensive theoretical and experimental analysis of common panoramic projections—spherical, cylindrical, and cubic—demonstrating the superior efficacy of cylindrical projection. Our method improves the mean absolute error (MAE) by 18.8% on the outdoor dataset Deep360 and by 19.9% on the real dataset 3D60. This work offers practical insights for other tasks and real-world applications, establishing a new paradigm in omnidirectional depth estimation. The code is available at `https://github.com/Qjizhi/MCPDepth`.

1. Introduction

Depth estimation is a pivotal challenge in geometric computer vision, playing a critical role in 3D scene understanding and robotic perception. Despite substantial advancements achieved through convolutional neural networks (CNNs) in processing perspective images, the task of estimating omnidirectional depth remains particularly challenging due to the severe geometric distortions inherent in panoramic representations. Recent research has investigated both monocular [14, 45] and stereo [18, 19, 47] approaches, each presenting unique advantages and limitations. Methods that apply conventional CNNs to spherical projections [14, 45, 47] often struggle to effectively manage these distortions, while those that directly model spherical epipolar geometry [18] encounter significant computational complexities. Although some strategies have introduced customized convolutional techniques, such as deformable convolution [42], EquiConvs [8], and spherical convolution [19], their practical deployment on resource-constrained robotic platforms remains a formidable challenge [39, 55]. Additionally, the inherent ambiguities associated with single or dual-view depth estimation frequently result in unreliable outputs, further complicating the task.

Several works have explored multi-view approaches, such as SweepNet [53] and OmniMVS [52], which use fisheye cameras to capture a panoramic field of view (FoV). However, these methods face limitations, including ineffective feature extraction due to severe radial distortions and ultra-wide FoV when using standard 2D convolutions, and incomplete depth reconstruction due to blind spots in fisheye camera configurations, leading to discontinuities in the spherical cost volume representation.

Recent research has advanced stereo matching for depth prediction by leveraging epipolar constraints to reduce the search space and improve accuracy. Notable contributions include 360SD-Net [47] and MODE [19], which have addressed challenges in complex depth estimation scenarios. MODE, in particular, introduces a two-stage framework that utilizes Cassini projection [50] to simplify epipolar geometry, followed by multi-view depth map fusion to enhance robustness. While these methods achieve state-of-the-art results, they face computational bottlenecks, especially on resource-constrained devices [39], due to their reliance on spherical convolutions [6]. Additionally, Cassini projection introduces significant distortions, particularly near the poles, which can degrade depth map quality.

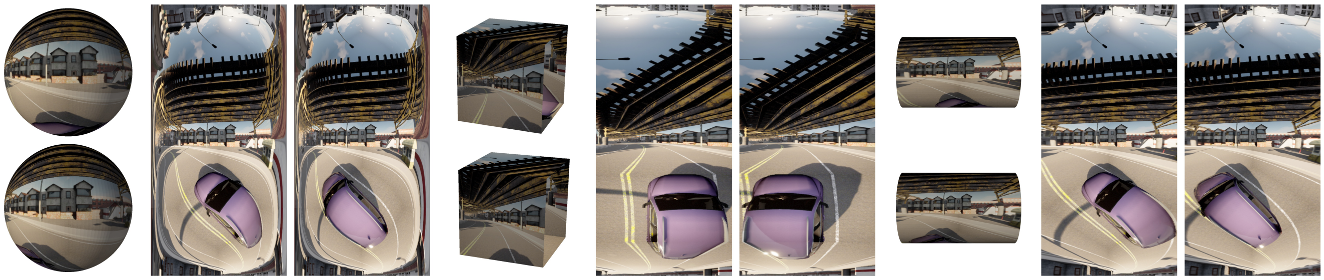

Figure 1. Comparison of stereo images among common panoramic projections.

Despite significant advancements in this field, the influence of projection methods on feature extraction and downstream tasks remains insufficiently explored. As illustrated in Figure 1, different projections exhibit distinct characteristics, each influencing the performance of CNNs and downstream tasks. In this work, we systematically analyze these effects and demonstrate that cylindrical projection is particularly effective for CNN-based feature extraction.

Drawing from the strengths of stereo matching and two-stage frameworks, we propose MCPDepth, a novel framework that leverages cylindrical projection for stereo matching. Our approach offers three key advantages: it significantly reduces geometric distortion compared to spherical projection, it's compatible with standard 2D convolutions, avoiding computationally intensive spherical convolutions, and it preserves the stereo matching relationship of perspective images, enabling better transfer learning from existing models. Additionally, we introduce a circular attention module that captures long-range dependencies across the full 360° vertical field of view (FoV) while mitigating projection-induced distortions. Our contributions can be summarized as follows:

- We introduce the first framework for omnidirectional depth estimation that leverages stereo matching across multiple cylindrical panoramas.
- We conduct a comprehensive theoretical and experimental analysis comparing common projections, highlighting the advantages of cylindrical projection.
- We present an innovative circular attention module designed to alleviate vertical axis distortions in cylindrical panoramas while significantly enhancing the receptive fields of conventional convolutions.
- Our method sets new benchmarks on the Deep360 (outdoor) and 3D60 (indoor) datasets.

2. Related Work

2.1. Deep Learning-based Stereo Matching

Early methods employed deep neural networks to compute matching costs, such as MCCNN [61], which trains a CNN for initial patch matching costs. Recently, end-to-end neural networks have dominated stereo matching methods. Works such as [10, 21, 22, 28, 30, 41, 57] only use 2D convolutions. Mayer *et al.* [28] propose the first end-to-end disparity estimation network, DispNet, and its correlation version, DispNetC. Pang *et al.* [30] introduce a two-stage framework named CRL with multi-scale residual learning. GwcNet [10] proposes the group-wise correlation volume to improve the expressiveness of the cost volume and performance in ambiguous regions.

AANet [57] adopts a novel aggregation algorithm using sparse points and multi-scale interaction. Another series of works [2, 15] use 3D convolutions, which demonstrate great potential in regularizing or filtering the cost volume. GCNet [15] first implements a 3D encoder-decoder architecture aimed at regularizing a 4D concatenation volume. PSMNet [2] proposes a stacked hourglass 3D CNN in conjunction with intermediate supervision to regularize the concatenation volume. Recently, iterative methods [17, 25, 44, 56] have shown impressive results. RAFTStereo [25] proposes to recurrently update the disparity field using local cost values retrieved from the all-pairs correlations. IGEV-Stereo [56] further advances this iterative approach by introducing a geometry encoding volume to encode non-local geometry and context information. Selective-Stereo [49] proposes a novel iterative update operator SRU for iterative stereo matching methods.

In parallel, substantial progress has been made in multi-view stereo (MVS) techniques [3, 9, 58, 59], which focus on generating 3D reconstructions from multiple perspective views, albeit primarily designed for limited-FoV cameras.

2.2. Omnidirectional Depth Estimation

Omnidirectional depth estimation has developed tremendously with neural networks. Zioulis *et al.* [65] present a learning-based monocular depth estimation method, trained directly on omnidirectional content in the ERP domain, and later propose CoordNet [66] with a spherical disparity model. BiFuse [45] uses both equirectangular and cubemap projections for depth estimation. A more effective fusion framework for ERP and cubemap projection is proposed in Unifuse [14]. Cheng *et al.* [4] introduce a depth sensing system by combining an OmniCamera with a regular depth sensor. 360SD-Net [47] is the first end-to-end trainable network for stereo depth estimation using spherical panoramas. CSDNet [18] focuses on left-right stereo and uses Mesh CNNs [13] to overcome spherical distortion. SweepNet [53] and OmniMVS [52] use multi-view fish-eye images for omnidirectional depth maps. However, most of them are based on spherical projection and extract spherical features with regular convolutions, and none of them discuss the properties of cylindrical projection.

Cheng *et al.* [4] propose a spherical feature transform layer to reduce the difficulty of feature learning. MODE [19] adopts spherical convolution from Spherenet [6], but the customized CUDA implementation poses deployment challenges on robotic platforms [39].

Jun *et al.* [38] employ cylindrical panoramas for stereo matching, but without CNNs or analysis of cylindrical projection properties, and 12 perspective images are stitched to obtain the panoramas.

2.3. Self-Attention Module

Attention mechanisms were first introduced by [1] for the encoder-decoder in a neural sequence-to-sequence model to capture token correspondence between sequences. Self-attention, designed for single contexts, encodes long-range interactions and has been widely applied in computer vision, achieving state-of-the-art performance [5, 12, 29, 31, 32, 40, 43, 48, 62]. Global self-attention in image processing is computationally expensive due to the need to calculate the relationship between every pixel and every other pixel, limiting its practical usage across all layers in a full-attention model. It is shown in [11, 35] that self-attention layers alone could form a fully attentional model by restricting the receptive field of self-attention to a local region.

In stereo matching, CREStereo [17] first adopts the self-attention module from LoFTR [40]. Zhao *et al.* [64] propose a multi-stage and multi-scale channel-attention transformer to preserve high-frequency information. GOAT [26] uses self-cross attention to capture more representative and distinguishable features. However, these methods are not designed for stereo matching in 360° panoramic images.

More recently, some attention mechanisms [24, 37, 60, 63] specifically designed for ERP have been proposed. However, these methods cannot be directly applied to cylindrical projection and face significant deployment challenges and computational overhead.

3. Method

Given m 360° cameras, where $m \geq 3$, we have a set of $n = \binom{m}{2} = \frac{m!}{2!(m-2)!}$ pairs of rectified panoramas $\{(I_L^i, I_R^i)\}_{i=1}^n$ with their intrinsic and extrinsic parameters. Our objective is to estimate the omnidirectional depth map d for the left panorama in the first pair I_L^1.

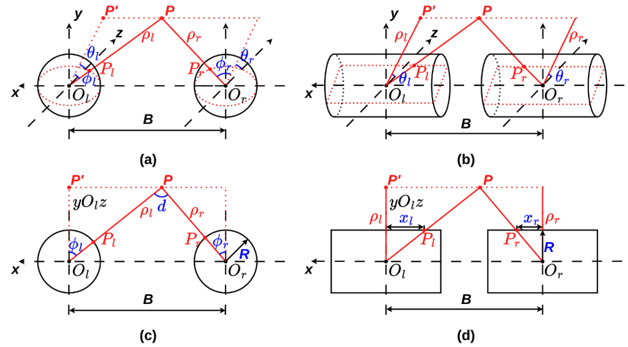

Figure 2. (a) and (b) compare the spherical and cylindrical projections for stereo matching and their respective epipolar geometries. (c) and (d) represent the schematic drawing of the epipolar plane under spherical and cylindrical projections.

3.1. Preliminaries: Panorama Projections

In this section, we discuss the similarities and differences between spherical and cylindrical projections for stereo matching. Here, cylindrical refers specifically to the vertical cylindrical projection. We omit details of the cubic projection, as it follows the same principles as perspective images. Finally, we provide a comprehensive analysis of the advantages and disadvantages of common panoramic projections.

As illustrated in Fig. 2, both cylindrical and spherical projections preserve the linear epipolar constraint. In spherical coordinates, ρ represents the Euclidean distance from the origin O to point P; ϕ is the angle between line OP and the plane yOz; and θ is the angle between line OP' and the z-axis, where P' is the projection of P on the plane yOz. In cylindrical coordinates, ρ denotes the Euclidean distance from the x-axis to point P; θ is the angle between line OP' and the z-axis, where P' is the projection of P on the plane yOz. The conversion between spherical, cylindrical, and Cartesian coordinate systems is illustrated in Eq. (1).

$$\begin{cases} x = \rho \sin(\phi) \\ y = \rho \cos(\phi) \sin(\theta) \\ z = \rho \cos(\phi) \cos(\theta) \end{cases} \quad \begin{cases} x = x \\ y = \rho \sin(\theta) \\ z = \rho \cos(\theta) \end{cases} \tag{1}$$

The spherical and cylindrical panoramas in Fig. 1 (b) and (f) are generated according to Eq. (2), where u and v are pixel coordinates, W and H are panorama dimensions and $R = H/2\pi$ is the cylinder's radius, which is the focal length in perspective images. u in the cylindrical panorama is the same as it is in the perspective images.

$$\begin{cases} u = (\phi + \frac{\pi}{2}) \cdot \frac{W}{\pi} \\ v = (\theta + \pi) \cdot \frac{H}{2\pi} \end{cases} \quad \begin{cases} u = -\frac{xR}{\rho} + \frac{W}{2} = -\frac{x}{\rho} \cdot \frac{H}{2\pi} + \frac{W}{2} \\ v = (\theta + \pi) \cdot \frac{H}{2\pi} \end{cases} \tag{2}$$

Table 1. Comparison of different projection types. h and v represent the horizontal and vertical FoV, respectively.

Projection Type	Advantages	Disadvantages
ERP	• Full coverage: $360°(h) \times 180°(v)$ FoV.	• Non-linear epipolar geometry, complicating stereo matching.
Cassini	• Linear epipolar geometry simplifies stereo matching. • Full coverage: $360°(v) \times 180°(h)$ FoV.	• Severe distortion near poles, uneven distortion. • Requires custom kernels for processing.
Cubic	• Linear epipolar geometry. • No distortion within individual cube faces. • Compatible with standard convolutional kernels.	• Limited horizontal FoV: $360°(v) \times 90°(h)$. • Discontinuities at cube joints, hindering CNN feature learning. • Requires fusion module across faces.
Cylindrical	• Linear epipolar geometry. • No distortion along the horizontal axis. • Uniform distortion along the vertical axis. • Compatible with standard convolutional kernels.	• Limited horizontal FoV: $360°(v) \times n°(h)$, where $n < 180$. • Residual distortion along the vertical axis.

In distortion-free perspective images, an object's actual length and its pixel length along the horizontal and vertical axes is given by $\Delta u = \frac{f_x}{z}\Delta x$ and $\Delta v = \frac{f_y}{z}\Delta y$, where f_x and f_y are the focal lengths along the x and y axes, and z is the distance along the z-axis. Eq. (3) shows these relationships for both spherical and cylindrical projections.

$$\begin{cases} \Delta u = f_\phi \Delta\phi \approx \frac{f_\phi}{\rho\cos\theta}\Delta X \\ \Delta v = f_\theta \Delta\theta \approx \frac{f_\theta}{\rho}\Delta Y \end{cases} \quad \begin{cases} \Delta u = \frac{f_X}{\rho}\Delta X \\ \Delta v = f_\theta \Delta\theta \approx \frac{f_\theta}{\rho}\Delta Y \end{cases} \tag{3}$$

where $f = R = H/2\pi$. The relation $\Delta u = f\Delta X/\rho$ and approximation $\Delta v \approx f\Delta Y/\rho$ for cylindrical projection hold under the condition that the object is not too large or far from the camera [33]. This approximation means objects in cylindrical projection appear similar regardless of their location. This **shift-invariant property** facilitates efficient learning by CNNs. In contrast, objects in spherical projection vary with their θ axis position, limiting the effectiveness of regular convolutions.

In addition, the disparity in spherical projection is defined as angular disparity d (Fig. 2 (c)), where $d = |\phi_l - \phi_r|$. This concept has been previously discussed in some works [19, 20, 23, 66]. The relationship between disparity and depth is:

$$\rho_l = B \cdot \frac{\sin(\phi_r + \frac{\pi}{2})}{\sin(d)} = B \cdot \left[\frac{\sin(\phi_l + \frac{\pi}{2})}{\tan(d)} - \cos(\phi_l + \frac{\pi}{2})\right] \tag{4}$$

where B denotes the baseline. As shown in Fig. 2 (d), the cylindrical projection maintains the same disparity-depth relationship as perspective images:

$$\rho_l = \frac{B \cdot f}{|x_l - x_r|} \tag{5}$$

Tab. 1 summarizes the advantages and disadvantages of different projection types. Cylindrical projection is the most suitable for stereo matching of panoramas due to the following reasons: (1) Compatibility with perspective images: Cylindrical panoramas maintain a disparity definition consistent with perspective images, enabling the direct application of stereo networks originally designed for perspective images. (2) Reduced distortion: Cylindrical panoramas only distort vertically, providing better shift invariance, which enhances CNN feature learning. (3) Simplified Deployment: Spherical panoramas require customized convolutions [6, 8, 42] to extract features. For example, spherical convolutions can't be exported to widely used ONNX [51] format for deployment, they either need CUDA Plugins for the TensorRT engine on NVIDIA platforms or customized implementation on other embedded devices. Cylindrical panoramas only use regular convolutions, making MCPDepth deployment-friendly.

3.2. Framework

The MCPDepth framework, shown in Fig. 3, includes two stages. In the stereo matching stage, n pairs of rectified cylindrical panoramas (Fig. 3 (a)) are fed into the stereo matching network. The number of pairs (n) and cameras (m) varies on different datasets: $n = 6, m = 4$ for Deep360, and $n = 3, m = 3$ for 3D60. The resulting disparity and confidence maps (Fig. 3 (b)) are reprojected into the Cassini domain with a 180° horizontal FoV. The disparity maps are then converted to depth maps. The depth and confidence maps are aligned with the view of I_L^1 using extrinsic parameters as shown in Fig. 3 (c). Black areas indicate invisible and occluded regions.

We use the circular attention module between feature extraction and cost volume with a structure similar to PSMNet [2]. The circular attention module augments the extracted features to capture features from a 360° FoV and overcome vertical-axis distortion. These augmented features are then shifted and concatenated to build the cost volume. The disparity map is regressed through the 3D stacked hourglass network. During training, we use the ℓ_1 loss to train the network. The confidence maps are used to measure the reliability of the disparity estimation and are widely used in stereo matching tasks [34]. The confidence map is obtained during inference. Specifically, considering the disparity is obtained through a probability-weighted sum over all disparity hypotheses, we compute the corresponding confidence value by taking a sum of probabilities over the three nearest disparity hypotheses.

We generally follow MODE's depth fusion stage structure. Specifically, multi-view depth maps, along with their corresponding confidence maps and reference panoramas,

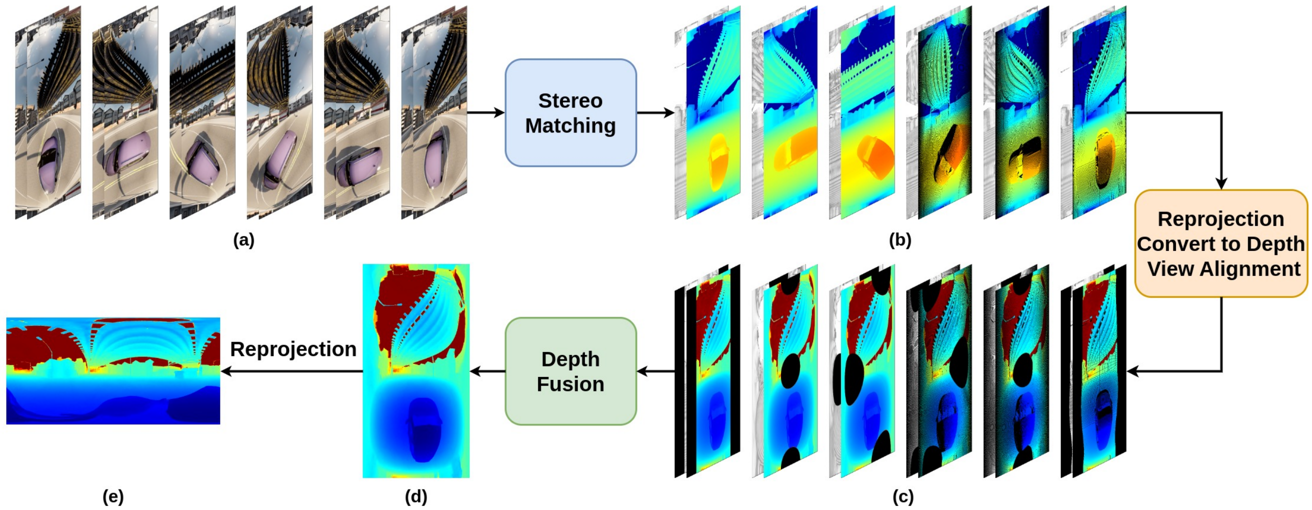

Figure 3. Framework of MCPDepth. (a) represents 6 pairs of cylindrical panoramas, (b) shows the disparity and confidence maps, and (c) shows the depth and confidence maps. (d) and (e) illustrate the depth map in Cassini and spherical projection.

are fed into two separate 2D encoder blocks. The fused depth map is then processed through a single decoder block, incorporating skip connections between the encoder and decoder blocks at each scale. The final depth map is generated in the Cassini domain [50] as shown in Fig. 3 (d), a transverse variant of the equirectangular projection (ERP) commonly used in map projections Fig. 3 (e), but it can be readily converted to the ERP domain. More details are available in the supplementary material.

3.3. Circular Attention

To overcome vertical axis distortion and capture the circular 360° features, we introduce a circular attention module. Conventional CNNs have limited receptive fields, which is restrictive for 360° FoV panoramas. The circular attention module, placed between feature extraction and cost volume construction is flexible and can be easily integrated since it maintains the input dimension. Besides, it only calculates the relations along the vertical axis, conserving more computing costs compared to global self-attention approaches. Fig. 4 (a) demonstrates our circular attention module.

In global self-attention, given an input feature map $x \in \mathbb{R}^{h\times w\times d_{in}}$ with height h, width w, and channels d_{in}. The output $y_o \in \mathbb{R}^{d_{out}}$ at position $o = (i, j)$ can be calculated as:

$$y_o = \sum_{p\in\mathcal{N}} softmax_p(q_o^T k_p) v_p \tag{6}$$

where $\mathcal{N}$ is the whole location lattice, $p = (a, b)$ are all possible positions. Queries $q_o = W_Q x_o$, keys $k_o = W_K x_o$, and values $v_o = W_V x_o$ are all linear projections of the input x_o, where $\forall o \in \mathcal{N}$. $W_Q, W_K \in \mathbb{R}^{d_q\times d_{in}}$, and $W_V \in \mathbb{R}^{d_{out}\times d_{in}}$ are all learnable weights.

However, global self-attention is extremely resource-consuming and computes $(\mathcal{O}(h^2w^2))$. Inspired by [11, 35], we restrict the receptive field of self-attention to a local region and apply only along the vertical axis. Additionally, global self-attention doesn't contain positional information, which is proven to be effective in many works [35, 36, 46, 54]. We incorporate positional information in the circular attention module. The output y_o at position $o = (i, j)$ can be calculated as:

$$y_o = \sum_{p\in\mathcal{N}_{1\times m}(o)} softmax_p(q_o^T k_p + q_o^T r^q_{p-o} + k_p^T r^k_{p-o})(v_p + r^v_{p-o}) \tag{7}$$

where $\mathcal{N}_{1\times m}(o)$ is the local $1 \times m$ region centered around location $o = (i, j)$. $r^q_{p-o} \in \mathbb{R}^{d_q}$ is the learnable relative positional encoding for queries and the inner product $q_o^T r^q_{p-o}$ measures the compatibility from location p to location o. Similarly, the learnable vectors $r^k_{p-o} \in \mathbb{R}^{d_q}$ and $r^v_{p-o} \in \mathbb{R}^{d_{out}}$ are positional encodings for keys and values. Our circular attention reduces the computation to $(\mathcal{O}(hwm))$.

For the Deep360 dataset, the feature map size after feature extraction is $h \times w \times d_{in} = 256 \times 128 \times 32$. After a 1×1 convolution is applied, the feature map is fed into a multi-head attention module, where the attention mechanism is only applied along the vertical axis. We set span $m = 256$ to ensure it captures all features along the vertical axis. We use 8 heads, each producing $256\times128\times4$ outputs. These are concatenated to $256\times128\times32$, and after another 1×1 convolution, the feature map is added element-wise to the original. Fig. 4 (b) illustrates how one head of the circular attention module works.

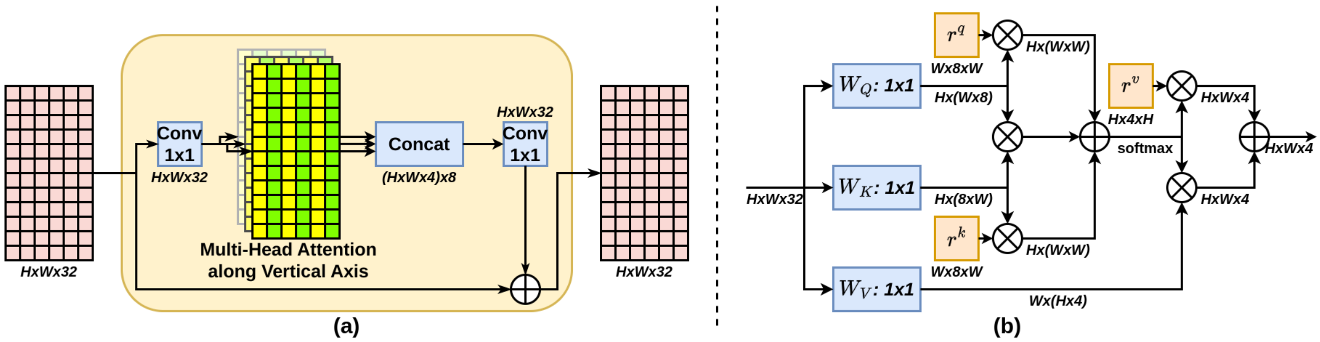

Figure 4. (a) displays the circular attention module in the stereo matching network. (b) represents our attention applied along the vertical axis. $\bigoplus$ denotes element-wise sum. $\bigotimes$ denotes matrix multiplication. Blue boxes are 1×1 convolution and orange boxes are relative positional encoding.

4. Experiments

4.1. Datasets

We train and evaluate our framework on Deep360 [19] and 3D60 [65], which include outdoor and indoor scenes. We evaluate both stereo matching and depth estimation. For Deep360, four 360° cameras are arranged horizontally in a square. Panoramas from all four views are used for evaluation. We use six stereo pairs for training and testing. For 3D60, three 360° cameras are arranged vertically in an equilateral right triangle. Panoramas from two of three views are used for evaluation. The resolutions are 1024×512 and 512×256, respectively.

4.2. Evaluation Metrics

Following MODE [19], we evaluate stereo matching performance using MAE (mean absolute error), RMSE (root mean square error), Px1,3,5 (percentage of outliers with pixel error $>$ 1, 3, 5), D1 [30] (percentage of outliers with pixel error $>$ 3 and $>$ 5%). We evaluate depth estimation performance using MAE, RMSE, AbsRel (absolute relative error), SqRel (square relative error), SILog [7](scale-invariant logarithmic error), $\delta 1, 2, 3$ [16] (accuracy with threshold that $\max(\frac{\hat{y}}{y^\star}, \frac{y^\star}{\hat{y}}) < 1.25, 1.25^2, 1.25^3$).

4.3. Implementation Details

We apply nearest-neighbor interpolation for cylindrical/cubic disparity maps generalization and bilinear interpolation for cylindrical/cubic panoramas generalization, both derived from spherical inputs.

In the stereo matching stage, cylindrical panoramas have a 360° vertical FoV and a horizontal FoV of less than 180°. We evaluate the central part of disparity maps in the Cassini domain with horizontal FoV = $2\arctan(\pi/2) \approx 105°$ for both datasets. This FoV yields cylindrical panorama size equivalent to spherical. In the fusion stage, we evaluate the entire omnidirectional depth map with a 360° horizontal and 180° vertical FoV.

Table 2. Quantitative results of stereo matching models pre-trained on perspective datasets evaluated on the Deep360 test dataset under different projections.

Method	Projection	MAE	Px1 (%)	D1 (%)
PSMNet [2]	Cassini	2.7667	42.7912	12.6288
	Cylindrical	**2.6118**	**34.4403**	**10.8204**
IGEV-Stereo [56]	Cassini	6.5155	61.0948	29.7265
	Cylindrical	**4.0194**	**53.3429**	**22.8117**
CREStereo [17]	Cassini	4.6836	43.5014	18.5130
	Cylindrical	**2.1241**	**22.6015**	**11.2502**

4.4. Experimental Results

Training on Perspective Images and Testing on Panoramas The pre-trained models of PSMNet [2] and IGEV-Stereo [56] are trained on Scene Flow [27], which contain only perspective images. CREStereo [17], trained on mixed datasets, exhibits better generalization. The performance of stereo matching with different projections on Deep360 is shown in Tab. 2. Acquiring panoramas and their depth ground truth is difficult, the experimental results demonstrate the potential to apply stereo-matching models trained on perspective images to cylindrical panoramas.

Comparisons with State-of-the-Art Methods We first evaluate our method against leading stereo matching networks such as PSMNet [2], AANet [57], and 360SD-Net [47], which is designed for 360° stereo. We train these models on the Deep360 training dataset from scratch and test them on the Deep360 test dataset following the default experimental settings. Tab. 3 shows that our method achieves state-of-the-art performance.

For omnidirectional depth estimation, we compare our method with other multi-view omnidirectional depth estimation methods including UNiFuse [14], CSDNet [18], 360SD-Net [47], OmniMVS [52], and MODE [19]. We report the results from MODE. Tab. 4 shows that our method

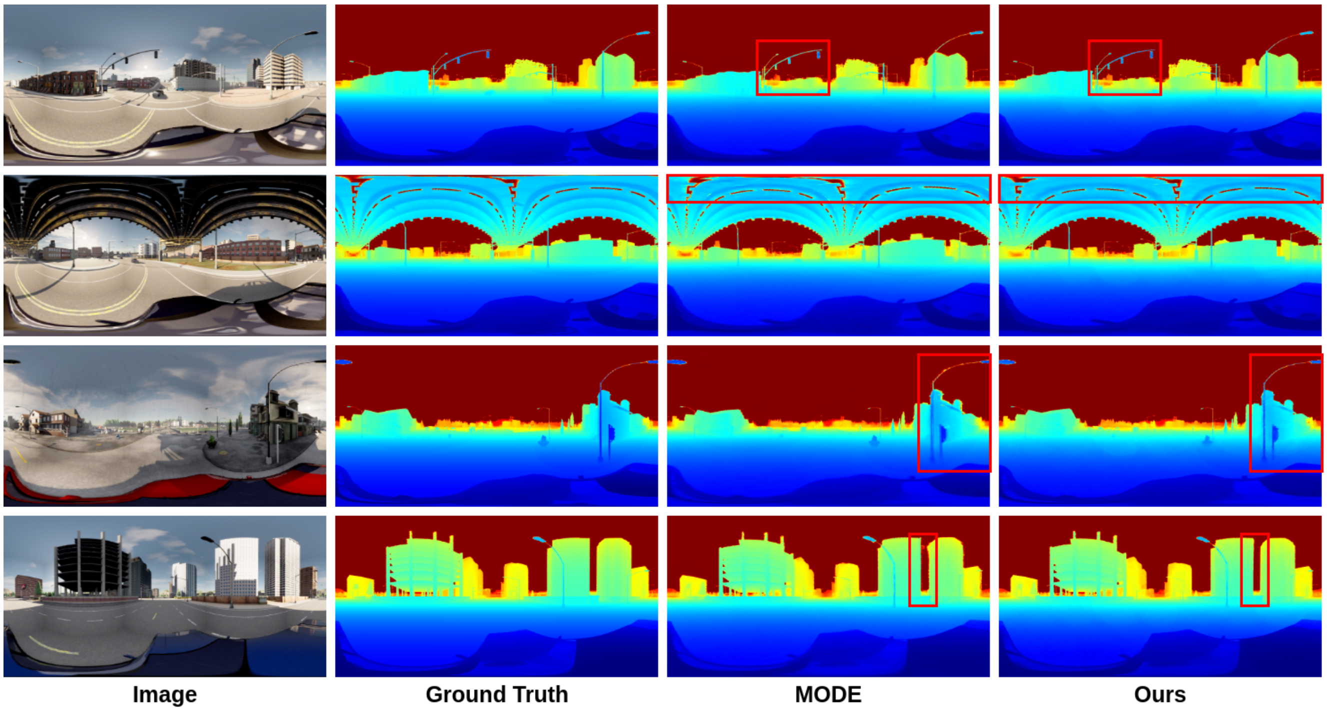

Figure 5. Depth estimation results on the Deep360 test dataset.

achieves an 18.8% MAE reduction on Deep360 and 19.9% on 3D60 compared to the previous best results, confirming its effectiveness for diverse real-world panoramas.

Fig. 5 illustrates the superior performance on Deep360, effectively handling severe distortions while preserving finer object details and the edges between the foreground and background better than MODE.

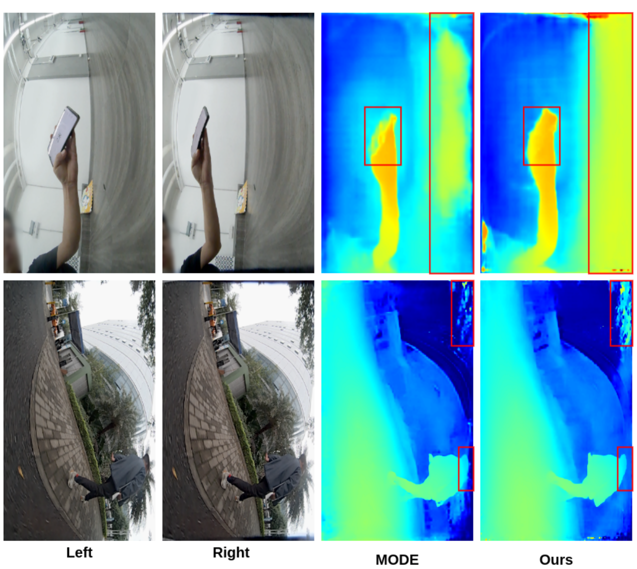

Figure 6. Depth estimation results in real-world scenarios.

Performance on Real Scenarios We evaluate our models on real-world fisheye image pairs. We reproject the fisheye images in Cassini projection, which have a vertical FoV of 189° and a horizontal FoV of 120°. The lens projection is equidistant, and we use OpenCV with checkerboards to calibrate the camera parameters and the relative pose of the two cameras. Fig. 6 shows the qualitative results of our models compared to MODE. For both indoor and outdoor scenes, with models trained on 3D60 and Deep360 respectively, MCPDepth demonstrates noticeable improvements over MODE, particularly in the highly distorted areas.

4.5. Ablation Study

Panorama Projection Tab. 5 demonstrates that cylindrical projection significantly outperforms spherical projection in stereo matching, even when applying spherical convolutions on spherical panoramas (MODE). Furthermore, we compare the performance of different projections on depth estimation in Tab. 4, demonstrating that cylindrical projection is the most suitable projection for regular convolutions, making it more effective for panorama stereo matching and depth estimation. These benefits may extend to other panorama-related vision tasks.

Circular Attention Tab. 6 shows that, although designed to mitigate vertical distortion for cylindrical projection, our circular attention module consistently improves performance across various panoramic projections and stereo-matching networks. This lightweight module delivers

Table 3. Quantitative results of stereo matching methods on Deep360 and 3D60 test datasets. The top three results for each metric are highlighted with a **first**, **second**, and **third** background, respectively.

Dataset	Method	Projection	Kernel Type	MAE ↓	RMSE ↓	Px1 (%) ↓	Px3 (%) ↓	Px5 (%) ↓	D1 (%) ↓
Deep360	AANet [57]	Cassini	Regular	0.3427	1.5703	5.2050	2.1515	1.2847	1.9817
	360SD-Net [47]	Cassini	Regular	0.5262	1.6459	3.8794	1.3389	0.8425	1.2989
	PSMNet [2]	Cassini	Regular	**0.2703**	**1.4790**	**3.3556**	**1.1979**	**0.7538**	**1.1708**
	MODE [19]	Cassini	Spherical	**0.2309**	**1.4014**	**2.8801**	**1.0488**	**0.6562**	**1.0326**
	Ours	Cylindrical	Regular	**0.2112**	**1.3903**	**2.5713**	**1.0009**	**0.6376**	**0.9828**
3D60	MODE [19]	Cassini	Spherical	0.2258	0.5265	2.9441	0.6482	0.2978	0.6478
	Ours	Cylindrical	Regular	**0.1773**	**0.4654**	**2.2298**	**0.5282**	**0.2564**	**0.5279**

Table 4. Quantitative results of omnidirectional depth estimation methods on Deep360 and 3D60 test datasets.

Dataset	Method	Kernel Type	MAE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	SILog ↓	$\delta 1\%$ ↑	$\delta 2\%$ ↑	$\delta 3\%$ ↑
Deep360	OmniMVS [52]	Regular	8.8865	59.3043	0.1073	2.9071	0.2434	94.9611	97.5495	98.2851
	360SD-Net [47]	Regular	11.2643	66.5789	0.0609	0.5973	0.2438	94.8594	97.2050	98.1038
	CSDNet [18]	Spherical	6.6548	36.5526	0.1553	1.7898	0.2475	86.0836	95.1589	97.7562
	UniFuse [14]	Regular	**3.9193**	**28.8475**	**0.0546**	**0.3125**	**0.1508**	**96.0269**	**98.2679**	**98.9909**
	MODE [19]	Spherical	**3.2483**	**24.9391**	**0.0365**	**0.0789**	**0.1104**	**97.9636**	**99.0987**	**99.4683**
	Ours+Cubic	Regular	5.0309	36.1907	0.0785	0.4410	0.1781	94.5960	98.1782	98.9406
	Ours	Regular	**2.6384**	**21.6692**	**0.0304**	**0.1153**	**0.1033**	**98.2557**	**99.2101**	**99.5227**
3D60	360SD-Net [47]	Regular	0.0762	0.2639	0.0300	**0.0117**	1.4578	97.6751	98.6603	99.0417
	CSDNet [18]	Spherical	0.2067	0.4225	0.0908	0.0241	0.1273	91.9537	98.3936	99.5109
	UniFuse [14]	Regular	**0.1868**	**0.3947**	**0.0799**	0.0246	**0.1126**	**93.2860**	**98.4839**	**99.4828**
	MODE [19]	Spherical	**0.0713**	**0.2631**	**0.0224**	**0.0031**	**0.0512**	**99.1283**	**99.7847**	**99.9250**
	Ours	Regular	**0.0571**	**0.1903**	**0.0199**	**0.0027**	**0.0401**	**99.3933**	**99.8506**	**99.9418**

significant accuracy gains with minimal additional computation, evaluated in the Cassini domain for spherical/cylindrical panoramas and the cubic domain for cubic panoramas. For IGEV-Stereo [56], applying circular attention to the largest feature map (first scale) yields substantial performance improvements.

Table 5. Ablation study for different projections on the Deep360 test dataset. The metrics refer to disparity errors.

Method	Projection	MAE	Px1 (%)	D1 (%)
MODE [19]	Cassini	0.2309	2.8801	1.0326
PSMNet [2]	Cassini	0.2703	3.3556	1.1708
	Cylindrical	**0.2179**	**2.6489**	**1.0236**
IGEV-Stereo [56]	Cassini	0.3905	6.1733	1.8843
	Cylindrical	**0.3278**	**4.7958**	**1.7276**

Table 6. Ablation study for circular attention module on the Deep360 test dataset. "CA" denotes circular attention. The metrics refer to disparity errors.

Method	Projection	CA	MAE	Px1 (%)	D1 (%)
MODE [19]	Cassini		0.2309	2.8801	1.0326
	Cassini	✓	**0.2210**	**2.7537**	**0.9881**
PSMNet [2]	Cubic		0.4471	5.0001	1.7623
	Cubic	✓	**0.4196**	**4.6699**	**1.6464**
PSMNet [2]	Cylindrical		0.2179	2.6489	1.0236
	Cylindrical	✓	**0.2112**	**2.5713**	**0.9828**
IGEV-Stereo [56]	Cylindrical		0.3278	4.7958	1.7276
	Cylindrical	✓	**0.2265**	**2.9581**	**1.1052**

5. Conclusion

We present MCPDepth, a novel two-stage framework for omnidirectional depth estimation through stereo matching from multiple cylindrical panoramas. Our comprehensive theoretical and experimental comparisons on different panoramic projections highlight the distinct advantages of cylindrical projection. Cylindrical projection maintains the linear epipolar constraint and preserves the definition of disparity as in perspective images. It effectively reduces distortion, enabling the application of stereo-matching models trained on perspective images to cylindrical panoramas. Additionally, cylindrical projection eliminates the need for customized kernels, simplifying deployment on embedded devices. Our circular attention module addresses vertical-axis distortions in cylindrical panoramas and captures 360° features, and can be extended to other projections. Experimental results demonstrate that MCPDepth achieves state-of-the-art performance on both the outdoor dataset Deep360 and the indoor dataset 3D60.

Limitations Cylindrical panoramas are limited in their ability to capture a full 180° horizontal FoV. As a result, at least 3 cameras are required to ensure complete coverage. Future work should explore optimizing the horizontal FoV of cylindrical panoramas to balance performance and computational resources.

Acknowledgments

We gratefully acknowledge the advanced computational resources provided by Engineering IT and Research Infrastructure Services at Washington University in St. Louis.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 3

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 2, 4, 6, 8

[3] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1538–1547, 2019. 2

[4] Xinjing Cheng, Peng Wang, Yanqi Zhou, Chenye Guan, and Ruigang Yang. Omnidirectional depth extension networks. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 589–595. IEEE, 2020. 3

[5] Peishan Cong, Xinge Zhu, Feng Qiao, Yiming Ren, Xidong Peng, Yuenan Hou, Lan Xu, Ruigang Yang, Dinesh Manocha, and Yuexin Ma. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19608–19617, 2022. 3

[6] 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)*, pages 518–533, 2018. 1, 3, 4

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 6

[8] Clara Fernandez-Labrador, Jose M Facil, Alejandro Perez-Yus, Cédric Demonceaux, Javier Civera, and Jose J Guerrero. Corners for layout: End-to-end layout recovery from 360 images. *IEEE Robotics and Automation Letters*, 5(2): 1255–1262, 2020. 1, 4

[9] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 2

[10] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3273–3282, 2019. 2

[11] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473, 2019. 3, 5

[12] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 3

[13] Chiyu Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al. Spherical cnns on unstructured grids. *arXiv preprint arXiv:1901.02039*, 2019. 3

[14] 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. 1, 3, 6, 8

[15] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 2

[16] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–96, 2014. 6

[17] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022. 2, 3, 6

[18] Ming Li, Xuejiao Hu, Jingzhao Dai, Yang Li, and Sidan Du. Omnidirectional stereo depth estimation based on spherical deep network. *Image and Vision Computing*, 114:104264, 2021. 1, 3, 6, 8

[19] Ming Li, Xueqian Jin, Xuejiao Hu, Jingzhao Dai, Sidan Du, and Yang Li. Mode: Multi-view omnidirectional depth estimation with 360○ cameras. In *European Conference on Computer Vision*, pages 197–213. Springer, 2022. 1, 3, 4, 6, 8

[20] Shigang Li. Binocular spherical stereo. *IEEE Transactions on intelligent transportation systems*, 9(4):589–600, 2008. 4

[21] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021. 2

[22] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2811–2820, 2018. 2

[23] Kaiwen Lin and Toby P Breckon. Real-time low-cost omnidirectional stereo vision via bi-polar spherical cameras. In *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*, pages 315–325. Springer, 2018. 4

[24] Zhixin Ling, Zhen Xing, Xiangdong Zhou, Manliang Cao, and Guichun Zhou. Panoswin: a pano-style swin transformer for panorama understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2023. 3

[25] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 2

[26] Zihua Liu, Yizhou Li, and Masatoshi Okutomi. Global occlusion-aware transformer for robust stereo matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3535–3544, 2024. 3

[27] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134. 6

[28] Nikolaus Mayer, Eddy Ilg, Philipp Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 2

[29] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2906–2917, 2021. 3

[30] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 887–895, 2017. 2, 6

[31] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018. 3

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 3

[33] Elad Plaut, Erez Ben Yaacov, and Bat El Shlomo. 3d object detection from a single fisheye image without a single fisheye training image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3659–3667, 2021. 4

[34] Matteo Poggi, Seungryong Kim, Fabio Tosi, Sunok Kim, Filippo Aleotti, Dongbo Min, Kwanghoon Sohn, and Stefano Mattoccia. On the confidence of stereo matching in a deep-learning era: a quantitative evaluation. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5293–5313, 2021. 4

[35] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in neural information processing systems*, 32, 2019. 3, 5

[36] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 5

[37] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: panorama transformer for indoor 360◦ depth estimation. In *European Conference on Computer Vision*, pages 195–211. Springer, 2022. 3

[38] Jun Shimamura, Naokazu Yokoya, Haruo Takemura, and Kazumasa Yamazawa. Construction of an immersive mixed environment using an omnidirectional stereo image sensor. In *Proceedings IEEE Workshop on Omnidirectional Vision (Cat. No. PR00704)*, pages 62–69. IEEE, 2000. 3

[39] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360 imagery. *Advances in neural information processing systems*, 30, 2017. 1, 3

[40] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 3

[41] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14372, 2021. 2

[42] 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)*, pages 707–722, 2018. 1, 4

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[44] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Itermvs: Iterative probability estimation for efficient multi-view stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8606–8615, 2022. 2

[45] 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*, pages 462–471, 2020. 1, 2

[46] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 5

[47] Ning-Hsu Wang, Bolivar Solarte, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. 360sd-net: 360 stereo depth estimation with learnable cost volume. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 582–588. IEEE, 2020. 1, 3, 6, 8

[48] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[49] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024. 2

[50] Wikipedia contributors. Cassini projection — Wikipedia, the free encyclopedia. https://en.wikipedia.org/

w/index.php?title=Cassini_projection&oldid=1184209037, 2023. [Online; accessed 29-February-2024]. 1, 5

[51] Wikipedia contributors. Open neural network exchange — Wikipedia, the free encyclopedia, 2023. [Online; accessed 6-March-2024]. 4

[52] Changhee Won, Jongbin Ryu, and Jongwoo Lim. Omnimvs: End-to-end learning for omnidirectional stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8987–8996, 2019. 1, 3, 6, 8

[53] Changhee Won, Jongbin Ryu, and Jongwoo Lim. Sweepnet: Wide-baseline omnidirectional depth estimation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6073–6079. IEEE, 2019. 1, 3

[54] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10041, 2021. 5

[55] Dawen Xu, Cheng Chu, Cheng Liu, Ying Wang, Huawei Li, Xiaowei Li, and Kwang-Ting Cheng. Energy-efficient accelerator design for deformable convolution networks. *arXiv preprint arXiv:2107.02547*, 2021. 1

[56] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 2, 6, 8

[57] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 2, 6, 8

[58] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020. 2

[59] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2

[60] 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*, pages 6101–6112, 2023. 3

[61] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1592–1599, 2015. 2

[62] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 3

[63] Junsong Zhang, Zisong Chen, Chunyu Lin, Zhijie Shen, Lang Nie, Kang Liao, and Yao Zhao. Sgformer: Spherical geometry transformer for 360 depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 3

[64] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1327–1336, 2023. 3

[65] 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)*, pages 448–465, 2018. 2, 6

[66] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, Federico Alvarez, and Petros Daras. Spherical view synthesis for self-supervised 360 depth estimation. In *2019 International Conference on 3D Vision (3DV)*, pages 690–699. IEEE, 2019. 2, 4