Novelty Assessment Report

Paper: DA$^2$: Depth Anything in Any Direction

PDF URL: https://openreview.net/pdf?id=323ximYcsk

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-05

Abstract

Panorama has a full FoV (360$^\circ\times$180$^\circ$), offering a more complete visual description than perspective images. Thanks to this characteristic, panoramic depth estimation is gaining increasing traction in 3D vision. However, due to the scarcity of panoramic data, previous methods are often restricted to in-domain settings, leading to poor zero-shot generalization. Furthermore, due to the spherical distortions inherent in panoramas, many approaches rely on perspective splitting (\textit{e.g.}, cubemaps), which leads to suboptimal efficiency. To address these challenges, we propose $\textbf{DA}$$^{\textbf{2}}$: $\textbf{D}$epth $\textbf{A}$nything in $\textbf{A}$ny $\textbf{D}$irection, an accurate, zero-shot generalizable, and fully end-to-end panoramic depth estimator. Specifically, for scaling up panoramic data, we introduce a data curation engine for generating high-quality panoramic depth data from perspective, and create $\sim$543K panoramic RGB-depth pairs, bringing the total to $\sim$607K. To further mitigate the spherical distortions, we present SphereViT, which explicitly leverages spherical coordinates to enforce the spherical geometric consistency in panoramic image features, yielding improved performance. A comprehensive benchmark on multiple datasets clearly demonstrates DA$^{2}$'s SoTA performance, with an average 38\% improvement on AbsRel over the strongest zero-shot baseline. Surprisingly, DA$^{2}$ even outperforms prior in-domain methods, highlighting its superior zero-shot generalization. Moreover, as an end-to-end solution, DA$^{2}$ exhibits much higher efficiency over fusion-based approaches. Both the code and the curated panoramic data will be released.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **panoramic depth estimation**

A total of **50 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Monocular Panoramic Depth Estimation**
- **Stereo and Multi-View Panoramic Depth Estimation**
- **Sensor Fusion and Depth Extension**
- **High-Resolution and Perspective-Panoramic Registration**
- **Application-Specific and Domain-Adapted Methods**
- **Datasets, Benchmarks, and Foundation Models**
- **Related Topics and Applications**

Complete Taxonomy Tree

- panoramic depth estimation Survey Taxonomy
- Monocular Panoramic Depth Estimation
    - Distortion-Aware Architectures
    - Deformable and Adaptive Convolution Methods (4 papers)
        - [11] Distortion-aware convolutional filters for dense prediction in panoramic images (Keisuke Tateno, 2018) View paper
        - [18] Omnidepth: Dense depth estimation for indoors spherical panoramas (N Zioulis, 2018) View paper
        - [25] Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation (Lu, 2022) View paper
        - [43] Distortion-aware monocular depth estimation for omnidirectional images (Chen Hong-xiang, 2021) View paper
    - Spherical Geometry and Coordinate-Based Methods ★ (4 papers)
        - [0] DA$^2$: Depth Anything in Any Direction (Anon et al., 2026) View paper
        - [12] Omnidirectional stereo depth estimation based on spherical deep network (Ming Li, 2021) View paper
        - [33] EGformer: Equirectangular Geometry-biased Transformer for 360 Depth Estimation (Ilwi Yun, 2023) View paper
        - [37] SPDET: Edge-Aware Self-Supervised Panoramic Depth Estimation Transformer With Spherical Geometry (Chuanqing Zhuang, 2023) View paper
    - Projection Fusion Architectures (5 papers)
        - [4] Unifuse: Unidirectional fusion for 360 panorama depth estimation (Hualie Jiang, 2021) View paper
        - [10] Omnifusion: 360 monocular depth estimation via geometry-aware fusion (Yuyan Li, 2022) View paper
        - [16] Bifuse: Monocular 360 depth estimation via bi-projection fusion (Fu-En Wang, 2020) View paper
        - [42] BiFuse++: Self-Supervised and Efficient Bi-Projection Fusion for 360Â° Depth Estimation (Wang, 2022) View paper
        - [49] HRDFuse: Monocular 360Â° Depth Estimation by Collaboratively Learning Holistic-with-Regional Depth Distributions (Hao Ai, 2023) View paper
    - Transformer-Based Architectures (4 papers)
    - [7] HiMODE: A Hybrid Monocular Omnidirectional Depth Estimation Model (Masum Shah Junayed, 2022) View paper
    - [15] PanoFormer: Panorama Transformer for Indoor 360Â° Depth Estimation (Shen, 2022) View paper
    - [19] GLPanoDepth: Global-to-Local Panoramic Depth Estimation (Jiayang Bai, 2022) View paper

  - [39] PCformer: A parallel convolutional transformer network for 360Â° depth estimation (Chao Xu, 2022) View paper
  - Multi-Task and Multi-Modal Learning (4 papers)
  - [9] Deep panoramic depth prediction and completion for indoor scenes (Giovanni Pintore, 2024) View paper
  - [13] MultiPanoWise: holistic deep architecture for multi-task dense prediction from a single panoramic image (Uzair Shah, 2024) View paper
  - [14] Sn360: Semantic and surface normal cascaded multi-task 360 monocular depth estimation (Payal Mohadikar, 2025) View paper
  - [27] Omnidirectional Depth Estimation for Semantic Segmentation (Jiaqi Zhou, 2024) View paper
  - Specialized Representation and Regularization Methods (5 papers)
  - [17] BGDNet: Background-guided Indoor Panorama Depth Estimation (Jiajing Chen, 2024) View paper
  - [20] Rethinking Supervised Depth Estimation for 360Â° Panoramic Imagery (Lu He, 2022) View paper
  - [28] Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation (G. Pintore, 2021) View paper
  - [30] Depth Estimation from Indoor Panoramas with Neural Scene Representation (Wenjie Chang, 2023) View paper
  - [50] Neural Contourlet Network for Monocular 360Â° Depth Estimation (Shen, 2022) View paper
  - Self-Supervised and Weakly-Supervised Methods (2 papers)
  - [26] PanoDepth: A Two-Stage Approach for Monocular Omnidirectional Depth Estimation (Li YuYan, 2021) View paper
  - [35] Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation (Yu-Lun Liu, 2024) View paper
- Stereo and Multi-View Panoramic Depth Estimation
  - Omnidirectional Stereo with Fisheye Cameras (4 papers)
  - [3] OmniStereo: Real-time Omnidireactional Depth Estimation with Multiview Fisheye Cameras (Jiaxi Deng, 2025) View paper
  - [6] OmniVidar: Omnidirectional Depth Estimation from Multi-Fisheye Images (Sheng Xie, 2023) View paper
  - [21] FastOmniMVS: Real-time Omnidirectional Depth Estimation from Multiview Fisheye Images (Yu-Shen Wang, 2023) View paper
  - [34] Depth Estimation using Omnidirectional Stereo Imaging and Machine Learning (Naoki Shibasaki, 2024) View paper
  - Equirectangular Stereo and Multi-View Fusion (3 papers)
  - [8] CasOmniMVS: Cascade Omnidirectional Depth Estimation with Dynamic Spherical Sweeping (Pinzhi Wang, 2024) View paper
  - [23] SDGE: Stereo Guided Depth Estimation for 360Â° Camera Sets (Jia-Lei Xu, 2024) View paper
  - [41] 360sd-net: 360 stereo depth estimation with learnable cost volume (Bolivar Solarte, 2020) View paper
- Sensor Fusion and Depth Extension (3 papers)
  - [2] Omnidirectional depth extension networks (Xinjing Cheng, 2020) View paper
  - [5] MODE: Monocular omnidirectional depth estimation via consistent depth fusion (Yunbiao Liu, 2023) View paper
  - [40] PanoFusion: A Monocular Omnidirectional Depth Estimation Model (Demidov, 2024) View paper
- High-Resolution and Perspective-Panoramic Registration (3 papers)
  - [24] High-resolution depth estimation for 360deg panoramas through perspective and panoramic depth images registration (CH Peng, 2023) View paper
  - [38] High-Resolution Depth Estimation for 360Â° Panoramas through Perspective and Panoramic Depth Images Registration (Chi-Han Peng, 2023) View paper
  - [46] High-resolution depth estimation for 360-degree panoramas through perspective and panoramic depth images registration (Peng, 2022) View paper
- Application-Specific and Domain-Adapted Methods (4 papers)
  - [22] Distortion-aware outdoor panoramic depth estimation via localâ□□global fusion (Ruyu Liu, 2025) View paper
  - [29] A novel panorama depth estimation framework for autonomous driving scenarios based on a vision transformer (Yuqi Zhang, 2024) View paper
  - [31] PanoDthNet: Depth Estimation Based on Indoor and Outdoor Panoramic Images (Jieyuan Cai, 2024) View paper
  - [48] Depth Estimation Using Single Fisheye Camera (Chun Kwok, 2023) View paper
- Datasets, Benchmarks, and Foundation Models (3 papers)
  - [1] Helvipad: A real-world dataset for omnidirectional stereo depth estimation (Mehdi Zayene, 2025) View paper
  - [32] Pano3d: A holistic benchmark and a solid baseline for 360deg depth estimation (G Albanis, 2021) View paper
  - [47] Depth Any Panoramas: A Foundation Model for Panoramic Depth Estimation (Xin Lin, 2025) View paper
- Related Topics and Applications (3 papers)
  - [36] Review on Panoramic Imaging and Its Applications in Scene Understanding (Shaohua Gao, 2022) View paper
  - [44] PanoVerse: automatic generation of stereoscopic environments from single indoor panoramic images for MetaverseÂ applications (Giovanni Pintore, 2023) View paper
  - [45] Rethinking supervised depth estimation for 360deg panoramic imagery (L He, 2022) View paper

Narrative

Core task: panoramic depth estimation. The field has evolved into several major branches that reflect different input modalities and architectural strategies. Monocular panoramic depth estimation forms the largest branch, encompassing distortion-aware architectures that handle spherical geometry through specialized convolutions, coordinate-based methods, and transformer-based designs. Stereo and multi-view approaches leverage multiple panoramic images to improve geometric consistency, while sensor fusion methods combine panoramic cameras with LiDAR or other modalities to extend depth range and accuracy. High-resolution and perspective-panoramic registration techniques address the challenge of aligning standard pinhole images with 360-degree views, and application-specific branches tailor depth estimation to domains such as autonomous driving or indoor scene understanding. Datasets, benchmarks, and foundation models provide the infrastructure for training and evaluation, with recent works like Depth Anywhere[35] and Depth Any Panoramas[47] exploring generalization across diverse panoramic scenarios.

Within monocular methods, a central tension exists between approaches that explicitly model spherical distortion versus those that adapt standard perspective architectures. Distortion-aware filters and spherical convolutions, exemplified by early work like Distortion Aware Filters[11] and Omnidepth[18], directly address equirectangular projection artifacts, while coordinate-based methods such as Spherical Deep Network[12] and EGformer[33] encode geometric priors through positional embeddings or spherical harmonics. The original paper, Depth Anything Direction[0], sits within this coordinate-based cluster, emphasizing directional cues in spherical space similarly to SPDET[37] and EGformer[33]. Compared to these neighbors, Depth Anything Direction[0] appears to push toward more flexible geometric representations that can generalize across viewing conditions, contrasting with SPDET[37]'s focus on explicit tangent-plane decompositions. This line of work reflects ongoing exploration of how best to encode 360-degree geometry without sacrificing the representational power of modern deep networks.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Omnidirectional stereo depth estimation based on spherical deep network

Authors: Ming Li, Xuejiao Hu, J. Dai, Yang Li, S. Du, et al. (7 authors total) | **Year/Venue**: 2021 | **URL**: View paper

Abstract

â¦ Omnidirectional depth estimation is an emerging research â¦ on sphere for omnidirectional depth estimation. We discuss the â¦ (SCRN) for omnidirectional depth estimation via the spherical â¦

Relationship Analysis

Both papers belong to the Spherical Geometry and Coordinate-Based Methods category, leveraging spherical coordinates and geometric constraints for panoramic depth estimation. They overlap in addressing spherical distortions through explicit spherical representations—DA^2 uses SphereViT with spherical embeddings in cross-attention, while the candidate employs a Spherical Convolution Residual Network (SCRN) for omnidirectional stereo depth estimation. The key difference is that DA^2 focuses on monocular zero-shot generalization with large-scale data curation, whereas the candidate appears to use stereo-based approaches with spherical convolutions for depth estimation.

2. EGformer: Equirectangular Geometry-biased Transformer for 360 Depth Estimation

Authors: Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, Chae Eun Rhee, et al. (6 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Estimating the depths of equirectangular (i.e., 360Â°) images (EIs) is challenging given the distorted 180Â° Ã 360Â° field-of-view, which is hard to be addressed via convolutional neural network (CNN). Although a transformer with global attention achieves significant improvements over CNN for EI depth estimation task, it is computationally inefficient, which raises the need for transformer with local attention. However, to apply local attention successfully for EIs, a specific strategy, which a...

Relationship Analysis

Both papers belong to the Spherical Geometry and Coordinate-Based Methods category, addressing panoramic depth estimation through explicit spherical geometric modeling. While DA^2 focuses on scaling up training data via a panoramic data curation engine and introduces SphereViT with spherical coordinate-based cross-attention for distortion-aware features, EGformer proposes equirectangular geometry-biased local attention mechanisms (ERPE, DAS, EaAR) within a hierarchical transformer architecture to handle distortions efficiently. The key difference is that DA^2 emphasizes zero-shot generalization through massive data scaling and end-to-end spherical embedding, whereas EGformer focuses on computationally efficient local attention with explicit geometric biases for in-domain performance.

3. SPDET: Edge-Aware Self-Supervised Panoramic Depth Estimation Transformer With Spherical Geometry

Authors: Chuanqing Zhuang, Zhengda Lu, Yiqun Wang, Jun Xiao, Ying Wang | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Panoramic depth estimation has become a hot topic in 3D reconstruction techniques with its omnidirectional spatial field of view. However, panoramic RGB-D datasets are difficult to obtain due to the lack of panoramic RGB-D cameras, thus limiting the practicality of supervised panoramic depth estimation. Self-supervised learning based on RGB stereo image pairs has the potential to overcome this limitation due to its low dependence on datasets. In this work, we propose the SPDET, an edge-aware sel...

Relationship Analysis

Both papers belong to the Spherical Geometry and Coordinate-Based Methods category, leveraging spherical coordinates and geometric constraints for panoramic depth estimation. They overlap in addressing spherical distortions through explicit geometric modeling—DA^2 uses spherical embeddings in cross-attention (SphereViT), while SPDET incorporates spherical geometry features into a transformer architecture. The key differences are: DA^2 focuses on zero-shot generalization through large-scale data curation (607K samples) and end-to-end learning, whereas SPDET emphasizes self-supervised learning from stereo pairs with edge-aware losses and depth-image-based rendering for novel view synthesis.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Panoramic data curation engine

Description: A pipeline that converts perspective RGB-depth pairs into full panoramic data through Perspective-to-Equirectangular projection and panoramic out-painting using FLUX-I2P. This engine scales up panoramic training data by approximately 10 times, significantly improving zero-shot generalization.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. 360 degree fish eye optical construction for equirectangular projection of panoramic images

URL: View paper

Brief Assessment

Fish Eye Optical[56] focuses on real-time fisheye-to-equirectangular projection for SLAM applications in autonomous driving, not on generating training data from perspective RGB-depth pairs for depth estimation models.

2. Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation

URL: View paper

Prior Art Analysis

Depth Anywhere[35] demonstrates prior work that converts perspective RGB-depth pairs into panoramic data through perspective-to-equirectangular projection, predating the ORIGINAL paper's claimed novelty. Both papers use P2E projection to map perspective images onto spherical space and employ panoramic out-painting techniques to generate complete panoramas. Depth Anywhere[35] explicitly describes using 'a six-face cube projection technique' and 'perspective depth estimation foundation models as teacher models to generate pseudo labels' for unlabeled 360-degree images, which parallels the ORIGINAL's approach of converting perspective data to panoramic format.

Evidence

Evidence 1 - **Rationale**: Both papers describe converting perspective images to panoramic format using projection techniques. The ORIGINAL uses P2E projection while Depth Anywhere[35] uses cube projection, but both address the same fundamental problem of generating panoramic depth data from perspective sources. - **Original**: given a perspective rgb image with known horizontal and vertical fovs, we first apply perspective-to-equirectangular (p2e) projection to map the image onto the spherical space. however, due to the limited fov of perspective images (with a typical horizontal range of 70 ◦-90◦), only a small portion o... - **Candidate**: our approach uses state-of-the-art perspective depth estimation models as teacher models to generate pseudo labels through a six-face cube projection technique, enabling efficient labeling of depth in 360-degree images

Evidence 2 - **Rationale**: Both papers describe generating complete panoramic depth data from perspective sources. The ORIGINAL uses panoramic out-painting with flux-i2p, while Depth Anywhere[35] uses perspective foundation models to generate pseudo labels, but both achieve the same goal of creating panoramic training data from perspective inputs. - **Original**: then, panoramic out-painting will be performed to generate a "complete" panorama to match the input of our model, using an image-to-panorama out-painter: flux-i2p (bfl, 2024; tencent, 2025). for the associated gt depth, we apply only the p2e projection without out-painting, due to concerns on the ab... - **Candidate**: our method leverages sota perspective depth estimation foundation models as teacher models and generates pseudo labels for unlabeled 360-degree images using a six-face cube projection approach. By doing so, we efficiently address the challenge of labeling depth in 360-degree imagery by leveraging pe...

Evidence 3 - **Rationale**: Both papers emphasize scaling up panoramic training data to improve model performance. The ORIGINAL claims this as a novel contribution, but Depth Anywhere[35] already described leveraging perspective data to augment 360-degree training datasets. - **Original**: this data curation engine substantially boosts the quantity and diversity of panoramic data, and significantly strengthens the zero-shot performance of da2, as shown in fig. 2 and tab. 2. - **Candidate**: To overcome these challenges, this paper presents a novel approach for training state-of-the-art (sota) depth estimation models on 360-degree imagery. With the recent significant increase in the amount of available data, the importance of both data quantity and quality has become evident.

3. Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image

URL: View paper

Brief Assessment

Deep Synthesis Omnidirectional[55] focuses on generating stereoscopic environments from panoramic images using depth estimation, not on converting perspective RGB-depth pairs to panoramic data through P2E projection and out-painting for training data augmentation.

4. Geometry-Aware Self-Supervised Indoor 360Â° Depth Estimation via Asymmetric Dual-Domain Collaborative Learning

URL: View paper

Brief Assessment

Asymmetric Dual Domain[52] mentions P2E transformation only in passing as a known technique. It does not describe a comprehensive data curation pipeline for converting perspective RGB-depth pairs into panoramic data, nor does it discuss panoramic out-painting or data scaling strategies that are central to the original contribution.

5. DA: Depth Anything in Any Direction

URL: View paper

Brief Assessment

Depth Anything Direction[51] describes the same data curation engine as the original paper, using identical P2E projection and FLUX-I2P out-painting methodology. This is the same work, not prior art that refutes novelty.

6. EpipolarGAN: Omnidirectional Image Synthesis with Explicit Camera Control

URL: View paper

Brief Assessment

EpipolarGAN[57] focuses on omnidirectional image synthesis with camera control, not on converting perspective RGB-depth pairs to panoramic data for depth estimation training. The technical approaches and objectives differ fundamentally.

7. High-resolution depth estimation for 360deg panoramas through perspective and panoramic depth images registration

URL: View paper

Brief Assessment

High Resolution Registration[24] focuses on stitching perspective depth maps to create high-resolution panoramic depth, not on converting perspective RGB-depth pairs into panoramic training data through out-painting for model training purposes.

8. DreamCube: RGB-D Panorama Generation via Multi-plane Synchronization

URL: View paper

Brief Assessment

DreamCube[53] focuses on RGB-D panorama generation using cubemap representations and multi-plane synchronization, not on converting perspective depth data to panoramic format through P2E projection and out-painting.

9. Unifuse: Unidirectional fusion for 360 panorama depth estimation

URL: View paper

Brief Assessment

Unifuse[4] focuses on fusing features from equirectangular and cubemap projections for depth estimation, not on generating panoramic training data from perspective images through P2E projection and out-painting.

10. Revisiting 360 Depth Estimation with PanoGabor: A New Fusion Perspective

URL: View paper

Brief Assessment

PanoGabor[54] does not address data generation or curation. It focuses on a Gabor-based fusion framework for handling distortions in existing 360° depth estimation, not on creating panoramic training data from perspective images.

Contribution 2: SphereViT architecture

Description: A Vision Transformer backbone that uses cross-attention with spherical embeddings derived from azimuth and polar angles. Image features attend to fixed spherical embeddings to produce distortion-aware representations, mitigating spherical distortions without requiring auxiliary modules or cubemap fusion.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Spherical Vision Transformers for Audio-Visual Saliency Prediction in 360-Degree Videos

URL: View paper

Brief Assessment

Spherical Vision Transformers[67] focuses on audio-visual saliency prediction in 360-degree videos using spherical geometry-aware attention layers, while the original paper addresses panoramic depth estimation with cross-attention mechanisms between image features and fixed spherical embeddings.

2. Mamba4PASS: Vision Mamba for PAnoramic Semantic Segmentation

URL: View paper

Brief Assessment

Mamba4PASS[68] uses Mamba-based architecture for panoramic semantic segmentation with a spherical geometry-aware deformable patch embedding module, which differs from the ORIGINAL paper's Vision Transformer backbone with cross-attention to fixed spherical embeddings for depth estimation.

3. Humanoidpano: Hybrid spherical panoramic-lidar cross-modal perception for humanoid robots

URL: View paper

Brief Assessment

HumanoidPano[69] focuses on humanoid robot perception using panoramic-lidar fusion for BEV semantic segmentation, not general panoramic depth estimation. While both use spherical embeddings, the candidate's application domain (robotic navigation) and task (semantic mapping) differ fundamentally from the original's depth estimation focus.

4. Distortion-aware outdoor panoramic depth estimation via localâ□□global fusion

URL: View paper

Brief Assessment

Local Global Fusion[22] addresses distortion through weight maps reflecting pixel-level distortion degrees in ERP images, not through cross-attention with spherical embeddings in a Vision Transformer architecture.

5. PanoFormer: Panorama Transformer for Indoor 360Â° Depth Estimation

URL: View paper

Prior Art Analysis

PanoFormer[15] demonstrates prior work that uses spherical coordinate embeddings in vision transformers for panoramic depth estimation. Both papers propose transformer architectures that leverage spherical coordinates (azimuth and polar angles) to create distortion-aware representations for panoramic images. PanoFormer[15] introduces a 'spherical token locating model (STLM)' that uses spherical coordinates to guide token sampling and position embedding, while the original paper proposes 'spherical embeddings' derived from azimuth and polar angles. Both approaches aim to mitigate spherical distortions without requiring cubemap fusion or auxiliary modules, using cross-attention mechanisms to inject spherical awareness into image features.

Evidence

Evidence 1 - **Rationale**: Both papers use spherical coordinates to create position embeddings that address distortions. PanoFormer[15]'s STLM uses spherical coordinates to locate tokens, similar to the original paper's spherical embeddings derived from azimuth and polar angles. - **Original**: we propose spherevit, which explicitly leverages spherical coordinates to enforce the spherical geometric consistency in panoramic image features, yielding improved performance - **Candidate**: we propose a distortion-based relative position embedding method in sec. 3.3. inspired by the cube projection, we note that the spherical tangent projection can effectively remove the distortion... we propose stlm to initialize the position of related tokens... the central token is projected from th...

Evidence 2 - **Rationale**: Both papers use attention mechanisms to inject spherical awareness. PanoFormer[15] uses a panorama self-attention mechanism with spherical token positions, while the original paper uses cross-attention with spherical embeddings as keys and values. - **Original**: spherevit uses cross-attention: image features are regarded as queries and the spherical embeddings as keys and values. this design lets the image feature explicitly attend to the panorama's spherical geometry, yielding distortion-aware representations and improved performance - **Candidate**: we redesign the self-attention module with additional learnable weight to push token flow, so as to flexibly capture various objects' structures... the psa can be represented as follows: psa(f, ŝs) = pm m=1 wm∗hphxw q=1 p9 k=1 amqk · w′ mf (ŝsmqk + Δsmqk)

Evidence 3 - **Rationale**: Both papers explicitly compute spherical coordinates (azimuth θ and polar ϕ angles) for each pixel position. PanoFormer[15] uses these coordinates to locate tokens on the spherical domain, demonstrating prior work using spherical coordinate embeddings. - **Original**: from the layout of erp, we first compute the spherical angles (azimuth and polar) of each pixel in the camera-centric spherical coordinates. after that, we expand this two-channel angle field into the image feature dimension using sine-cosine basis embedding, forming the spherical embedding - **Candidate**: let the unit sphere be s2, and s(0, 0) = (θ0, ϕ0) ∈ s2 is the spherical coordinate origin. ∀s(x, y) = (θ, ϕ) ∈ s2, we can obtain other 8 points (related tokens) around it (current token) on the spherical domain. s(±1, 0) =(θ ± Δθ, ϕ) s(0, ±1) =(θ, ϕ± Δϕ) s(±1, ±1) =(θ ± Δθ, ϕ± Δϕ)

Evidence 4 - **Rationale**: PanoFormer[15] was published earlier and proposes a transformer architecture with spherical coordinate-based position embeddings to handle panoramic distortions, demonstrating that the concept of using spherical embeddings in vision transformers for panoramic images existed before the original paper. - **Original**: to mitigate the impact of spherical distortion, inspired by the positional embeddings in vision transformers (vits), we propose spherevit-the main backbone of da2... this spherical embedding can be fixed and reusable - **Candidate**: we propose the first panorama transformer (panoformer) to enable the network's panoramic perception capability by removing distortions and perceiving geometric structures simultaneously... we design a relative position embedding method to reduce the negative effect of distortions, which utilizes a c...

6. SGAT4PASS: Spherical Geometry-Aware Transformer for PAnoramic Semantic Segmentation

URL: View paper

Brief Assessment

SGAT4PASS[70] focuses on panoramic semantic segmentation using spherical deformable patch embedding with intra- and inter-offset constraints, not on depth estimation. While both use spherical coordinates, SGAT4PASS[70] addresses segmentation tasks with different architectural goals than the depth-focused cross-attention mechanism in SphereViT.

7. A Comparison of Spherical Neural Networks for Surround-View Fisheye Image Semantic Segmentation

URL: View paper

Brief Assessment

Spherical Networks Comparison[71] evaluates existing spherical vision transformers on fisheye images for semantic segmentation, rather than proposing a novel architecture. The original paper introduces SphereViT with cross-attention mechanisms for panoramic depth estimation, which is a different task and architectural design.

8. SGFormer: Spherical Geometry Transformer for 360 Depth Estimation

URL: View paper

Brief Assessment

SGFormer[66] focuses on a spherical prior decoder with bipolar re-projection and circular rotation techniques, rather than cross-attention with spherical embeddings as queries/keys/values. The architectural approaches differ fundamentally in how spherical geometry is integrated into the transformer.

9. SphereUFormer: A U-Shaped Transformer for Spherical 360 Perception

URL: View paper

Brief Assessment

SphereUFormer[65] operates on spherical mesh representations (icospheres) using graph-based attention mechanisms, while the original paper's SphereViT uses cross-attention between image features and fixed spherical embeddings derived from azimuth/polar angles in equirectangular projection. These are fundamentally different architectural approaches to handling spherical distortions.

10. Spherical Vision Transformers for Audio-Visual Saliency Prediction in 360 Videos

URL: View paper

Brief Assessment

Spherical Vision Transformers[64] focuses on audio-visual saliency prediction in 360 videos, not panoramic depth estimation. The architectural approaches and application domains differ fundamentally from the original paper's depth estimation framework.

Contribution 3: Comprehensive benchmark for panoramic depth estimation

Description: A thorough evaluation framework comparing both zero-shot and in-domain methods, as well as panoramic and perspective approaches, across multiple recognized datasets. The benchmark demonstrates that DA2 achieves state-of-the-art zero-shot performance and even surpasses prior in-domain methods.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Depth Any Panoramas: A Foundation Model for Panoramic Depth Estimation

URL: View paper

Brief Assessment

Depth Any Panoramas[47] focuses on metric depth estimation with a different evaluation framework and does not challenge DA2's novelty in establishing a comprehensive benchmark comparing zero-shot/in-domain and panoramic/perspective methods.

2. Depth any camera: Zero-shot metric depth estimation from any camera

URL: View paper

Brief Assessment

Depth Any Camera[59] focuses on zero-shot metric depth estimation across different camera types (perspective, fisheye, 360°) using equirectangular projection, not specifically on panoramic depth estimation benchmarks. The candidate addresses camera generalization rather than comprehensive panoramic benchmarking methodologies.

3. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and MÃ¶bius Spatial Augmentation

URL: View paper

Brief Assessment

PanDA Mobius[62] focuses on semi-supervised learning with Möbius transformation augmentation for panoramic depth, not on establishing comprehensive benchmarks comparing zero-shot and in-domain methods across multiple datasets as claimed by the original paper.

4. DA: Depth Anything in Any Direction

URL: View paper

Brief Assessment

Depth Anything Direction[51] presents the same benchmark evaluation framework as the original paper, comparing zero-shot and in-domain methods across the same datasets. This is the same work, not prior art.

5. Open panoramic segmentation

URL: View paper

Brief Assessment

Open Panoramic Segmentation[61] focuses on panoramic semantic segmentation tasks with open-vocabulary learning, not depth estimation benchmarks. The paper addresses segmentation across different domains rather than zero-shot depth estimation evaluation frameworks.

6. Depth Anything in : Towards Scale Invariance in the Wild

URL: View paper

Brief Assessment

Scale Invariance Wild[63] focuses on zero-shot panoramic depth estimation with scale-invariant outputs and introduces a new outdoor benchmark (Metropolis). While both papers evaluate zero-shot methods, Scale Invariance Wild[63] does not challenge the novelty of creating a comprehensive benchmark framework that compares both zero-shot/in-domain and panoramic/perspective approaches across multiple datasets as claimed by the original paper.

7. Pano3d: A holistic benchmark and a solid baseline for 360deg depth estimation

URL: View paper

Brief Assessment

Pano3d[32] focuses on holistic evaluation methodology (direct depth, boundary preservation, smoothness) and cross-dataset generalization testing, while the original paper emphasizes zero-shot performance improvements through data curation and model architecture. These represent complementary rather than competing contributions to panoramic depth estimation benchmarking.

8. Sn360: Semantic and surface normal cascaded multi-task 360 monocular depth estimation

URL: View paper

Brief Assessment

Sn360[14] focuses on a multi-task cascaded framework for depth estimation using semantic and surface normal guidance, not on establishing comprehensive benchmarks comparing zero-shot versus in-domain methods across multiple datasets.

9. PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation

URL: View paper

Prior Art Analysis

PanDA[60] demonstrates that prior work has already established comprehensive benchmarking frameworks for panoramic depth estimation that compare zero-shot and in-domain methods across multiple datasets. The candidate paper presents extensive quantitative comparisons on the same well-recognized benchmarks (Matterport3D, Stanford2D3D) used by the original paper, evaluating both zero-shot methods (including perspective depth estimators adapted to panoramas) and in-domain panoramic methods. PanDA[60] explicitly states conducting 'comprehensive experiments to assess the performance' across multiple factors and provides detailed benchmark tables comparing various method categories, predating the original paper's claimed contribution of a 'comprehensive benchmark'.

Evidence

Evidence 1 - **Rationale**: Both papers claim to establish comprehensive benchmarks demonstrating state-of-the-art zero-shot performance that surpasses in-domain methods on recognized datasets, indicating PanDA[60] already established this type of benchmark. - **Original**: comprehensive benchmark on multiple datasets clearly demonstrates da2's sota performance, with an average 38% improvement on absrel over the strongest zero-shot baseline. surprisingly, da 2 even outperforms prior in-domain methods, highlighting its superior zero-shot generalization. - **Candidate**: extensive experiments demonstrate that panda exhibits remarkable zeroshot capability across diverse scenes, and outperforms the data-specific panoramic depth estimation methods on two popular real-world benchmarks.

Evidence 2 - **Rationale**: PanDA[60] explicitly describes conducting comprehensive empirical investigations across multiple critical factors for panoramic depth estimation, establishing a thorough evaluation framework before the original paper. - **Original**: to validate da2, we conduct a comprehensive benchmark on scale-invariant distance combining multiple well-recognized evaluation datasets. however, due to the scarcity of panoramic data, existing zero-shot approaches in panoramic depth estimation are limited, whereas in perspective, there exist many ... - **Candidate**: to this end, we conduct an empirical investigation into several critical factors that influence dams' performance on panoramas: 1) different representations of panoramas: the choice of representations is vital for the model to learn effective features. panoramas can be represented in various represe...

Evidence 3 - **Rationale**: Both papers present comprehensive benchmark tables comparing zero-shot methods (including perspective methods adapted to panoramas) across multiple datasets, with PanDA[60] establishing this comparison framework earlier. - **Original**: tab. 1 presents a comprehensive comparison of da 2 with previous sota approaches. following (wang et al., 2025c;d), we also include prior perspective methods for a more thorough comparison. as demonstrated in tab. 1, da 2 consistently outperforms all other methods across various settings. particular... - **Candidate**: as illustrated in tab. 4, we compare with zero-shot depth estimation methods designed for perspective images, e.g., dam v1 [47], dam v2 [48], and marigold [19]. the results demonstrate that our panda outperforms the other methods across all metrics and datasets, highlighting its effective zero-shot ...

Evidence 4 - **Rationale**: PanDA[60] explicitly describes benchmarking against both zero-shot foundation models and in-domain panoramic methods on the same standard datasets, demonstrating prior establishment of this comprehensive comparison framework. - **Original**: both zero-shot / in-domain, panoramic / perspective methods are compared to build a comprehensive benchmark for panoramic depth estimation. - **Candidate**: we leverage two real-world datasets-matterport3d [10] and stanford2d3d [5]-to access the zero-shot performance of panda in comparison with zero-shot depth foundation models. we also benchmark panda against sota panoramic depth estimation methods by fine-tuning it on real-world datasets.

10. Metric3d: Towards zero-shot metric 3d prediction from a single image

URL: View paper

Brief Assessment

Metric3d[58] focuses on zero-shot metric 3D prediction from single images across various camera types, not specifically on panoramic depth estimation benchmarks. The candidate paper's scope differs from DA2's comprehensive evaluation framework for panoramic methods.

Appendix: Text Similarity Detection

Textual similarity detection checked 32 papers and found 6 similarity segment(s) across 4 paper(s).

The following **4 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. DA: Depth Anything in Any Direction

Detected in: Contribution: contribution_1, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

2. Humanoidpano: Hybrid spherical panoramic-lidar cross-modal perception for humanoid robots

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

3. EGformer: Equirectangular Geometry-biased Transformer for 360 Depth Estimation

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

4. SGAT4PASS: Spherical Geometry-Aware Transformer for PAnoramic Semantic Segmentation

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] DA^2: Depth Anything in Any Direction View paper
- [1] Helvipad: A real-world dataset for omnidirectional stereo depth estimation View paper
- [2] Omnidirectional depth extension networks View paper
- [3] OmniStereo: Real-time Omnidireactional Depth Estimation with Multiview Fisheye Cameras View paper
- [4] Unifuse: Unidirectional fusion for 360 panorama depth estimation View paper
- [5] MODE: Monocular omnidirectional depth estimation via consistent depth fusion View paper
- [6] OmniVidar: Omnidirectional Depth Estimation from Multi-Fisheye Images View paper
- [7] HiMODE: A Hybrid Monocular Omnidirectional Depth Estimation Model View paper
- [8] CasOmniMVS: Cascade Omnidirectional Depth Estimation with Dynamic Spherical Sweeping View paper
- [9] Deep panoramic depth prediction and completion for indoor scenes View paper
- [10] Omnifusion: 360 monocular depth estimation via geometry-aware fusion View paper
- [11] Distortion-aware convolutional filters for dense prediction in panoramic images View paper
- [12] Omnidirectional stereo depth estimation based on spherical deep network View paper
- [13] MultiPanoWise: holistic deep architecture for multi-task dense prediction from a single panoramic image View paper
- [14] Sn360: Semantic and surface normal cascaded multi-task 360 monocular depth estimation View paper
- [15] PanoFormer: Panorama Transformer for Indoor 360Â° Depth Estimation View paper
- [16] Bifuse: Monocular 360 depth estimation via bi-projection fusion View paper
- [17] BGDNet: Background-guided Indoor Panorama Depth Estimation View paper
- [18] Omnidepth: Dense depth estimation for indoors spherical panoramas View paper
- [19] GLPanoDepth: Global-to-Local Panoramic Depth Estimation View paper
- [20] Rethinking Supervised Depth Estimation for 360Â° Panoramic Imagery View paper
- [21] FastOmniMVS: Real-time Omnidirectional Depth Estimation from Multiview Fisheye Images View paper
- [22] Distortion-aware outdoor panoramic depth estimation via localâ□□global fusion View paper
- [23] SDGE: Stereo Guided Depth Estimation for 360Â° Camera Sets View paper
- [24] High-resolution depth estimation for 360deg panoramas through perspective and panoramic depth images registration View paper
- [25] Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation View paper
- [26] PanoDepth: A Two-Stage Approach for Monocular Omnidirectional Depth Estimation View paper
- [27] Omnidirectional Depth Estimation for Semantic Segmentation View paper
- [28] Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation View paper
- [29] A novel panorama depth estimation framework for autonomous driving scenarios based on a vision transformer View paper
- [30] Depth Estimation from Indoor Panoramas with Neural Scene Representation View paper
- [31] PanoDthNet: Depth Estimation Based on Indoor and Outdoor Panoramic Images View paper
- [32] Pano3d: A holistic benchmark and a solid baseline for 360deg depth estimation View paper
- [33] EGformer: Equirectangular Geometry-biased Transformer for 360 Depth Estimation View paper
- [34] Depth Estimation using Omnidirectional Stereo Imaging and Machine Learning View paper
- [35] Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation View paper
- [36] Review on Panoramic Imaging and Its Applications in Scene Understanding View paper
- [37] SPDET: Edge-Aware Self-Supervised Panoramic Depth Estimation Transformer With Spherical Geometry View paper
- [38] High-Resolution Depth Estimation for 360Â° Panoramas through Perspective and Panoramic Depth Images Registration View paper
- [39] PCformer: A parallel convolutional transformer network for 360Â° depth estimation View paper
- [40] PanoFusion: A Monocular Omnidirectional Depth Estimation Model View paper
- [41] 360sd-net: 360 stereo depth estimation with learnable cost volume View paper
- [42] BiFuse++: Self-Supervised and Efficient Bi-Projection Fusion for 360Â° Depth Estimation View paper
- [43] Distortion-aware monocular depth estimation for omnidirectional images View paper
- [44] PanoVerse: automatic generation of stereoscopic environments from single indoor panoramic images for MetaverseÂ applications View paper
- [45] Rethinking supervised depth estimation for 360deg panoramic imagery View paper
- [46] High-resolution depth estimation for 360-degree panoramas through perspective and panoramic depth images registration View paper
- [47] Depth Any Panoramas: A Foundation Model for Panoramic Depth Estimation View paper
- [48] Depth Estimation Using Single Fisheye Camera View paper
- [49] HRDFuse: Monocular 360Â° Depth Estimation by Collaboratively Learning Holistic-with-Regional Depth Distributions View paper
- [50] Neural Contourlet Network for Monocular 360Â° Depth Estimation View paper
- [51] DA: Depth Anything in Any Direction View paper
- [52] Geometry-Aware Self-Supervised Indoor 360Â° Depth Estimation via Asymmetric Dual-Domain Collaborative Learning View paper
- [53] DreamCube: RGB-D Panorama Generation via Multi-plane Synchronization View paper
- [54] Revisiting 360 Depth Estimation with PanoGabor: A New Fusion Perspective View paper
- [55] Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image View paper
- [56] 360 degree fish eye optical construction for equirectangular projection of panoramic images View paper
- [57] EpipolarGAN: Omnidirectional Image Synthesis with Explicit Camera Control View paper
- [58] Metric3d: Towards zero-shot metric 3d prediction from a single image View paper
- [59] Depth any camera: Zero-shot metric depth estimation from any camera View paper

- [60] PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and Mobius Spatial Augmentation View paper
- [61] Open panoramic segmentation View paper
- [62] PanDA: Towards Panoramic Depth Anything with Unlabeled Panoramas and MÃ¶bius Spatial Augmentation View paper
- [63] Depth Anything in : Towards Scale Invariance in the Wild View paper
- [64] Spherical Vision Transformers for Audio-Visual Saliency Prediction in 360 Videos View paper
- [65] SphereUFormer: A U-Shaped Transformer for Spherical 360 Perception View paper
- [66] SGFormer: Spherical Geometry Transformer for 360 Depth Estimation View paper
- [67] Spherical Vision Transformers for Audio-Visual Saliency Prediction in 360-Degree Videos View paper
- [68] Mamba4PASS: Vision Mamba for PAnoramic Semantic Segmentation View paper
- [69] Humanoidpano: Hybrid spherical panoramic-lidar cross-modal perception for humanoid robots View paper
- [70] SGAT4PASS: Spherical Geometry-Aware Transformer for PAnoramic Semantic Segmentation View paper
- [71] A Comparison of Spherical Neural Networks for Surround-View Fisheye Image Semantic Segmentation View paper