RPG360: Robust 360 Depth Estimation with Perspective Foundation Models and Graph Optimization

Dongki Jung Jaehoon Choi Yonghan Lee Dinesh Manocha
University of Maryland, College Park
{jdk9405, kevchoi, lyhan12, dmanocha}@umd.edu

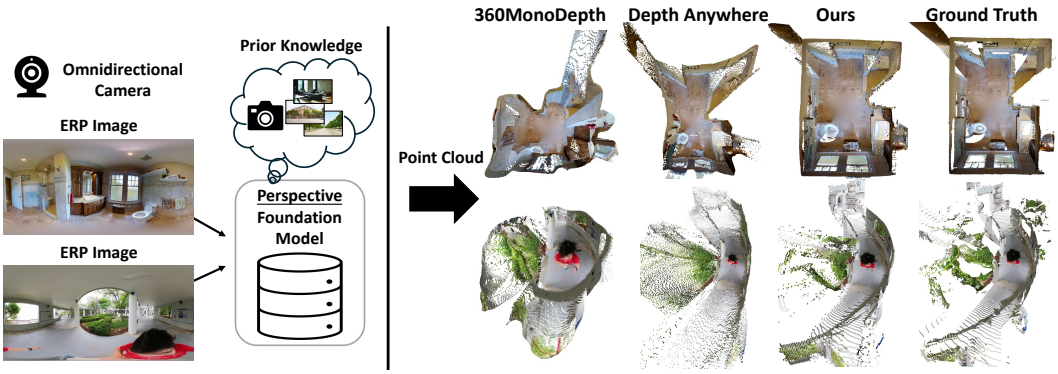

Figure 1: (Left) 360° images can be transformed into multiple undistorted tangential plane images, and then scale-ambiguous depth maps can be predicted using perspective-foundation models [46, 68, 25]. (Right) 360MonoDepth [47] and Depth Anywhere [60], have addressed the scale ambiguity issue using optimization or learning-based methods. However, these approaches suffer from 3D structural degradation, as seen in the reconstructed point cloud visualization. Our proposed method leverages the prior knowledge of the perspective foundation model along with graph optimization, thereby enhances 3D structural awareness and achieves superior performance.

Abstract

The increasing use of 360° images across various domains has emphasized the need for robust depth estimation techniques tailored for omnidirectional images. However, obtaining large-scale labeled datasets for 360° depth estimation remains a significant challenge. In this paper, we propose RPG360, a training-free **r**obust 360° monocular depth estimation method that leverages **p**erspective foundation models and **g**raph optimization. Our approach converts 360° images into six-face cubemap representations, where a perspective foundation model is employed to estimate depth and surface normals. To address depth scale inconsistencies across different faces of the cubemap, we introduce a novel depth scale alignment technique using graph-based optimization, which parameterizes the predicted depth and normal maps while incorporating an additional per-face scale parameter. This optimization ensures depth scale consistency across the six-face cubemap while preserving 3D structural integrity. Furthermore, as foundation models exhibit inherent robustness in zero-shot settings, our method achieves superior performance across diverse datasets, including Matterport3D, Stanford2D3D, and 360Loc. We also demonstrate the versatility of our depth estimation approach by validating its benefits in downstream tasks such as feature matching $3.2 \sim 5.4\%$ and Structure from Motion $0.2 \sim 9.7\%$ in AUC@5°. Project Page: https://jdk9405.github.io/RPG360/

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

360° sensors offer significant advantages due to their wide field of view, allowing the capture of rich contextual information within a single image [74, 9, 22, 66]. The growing adoption of such panoramic images in various applications, such as robot navigation [38, 65], virtual reality [35], autonomous vehicles [41], and immersive media [10], necessitates robust depth estimation, generalizing to zero-shot settings, techniques specifically tailored for 360° images. However, applying conventional perspective depth estimation methods [67, 68, 71, 25, 62] to 360° images directly is challenging due to fundamental differences in camera models, which introduce distortions. To address these challenges, previous research in 360° depth estimation has broadly followed two approaches: 1) learning-based methods and 2) optimization-based methods.

Existing *learning-based approaches* [78, 43, 73, 54, 27, 59, 2, 1, 77, 72, 28] have focused on designing 360° depth estimation networks trained on labeled 360° datasets. These methods attempt to mitigate distortions through adaptive network architectures [78, 53, 61], or by leveraging cubemap [58, 59, 27] and tangent image [37, 51] projections during training. However, a major limitation of these approaches is the scarcity of labeled datasets for 360° depth estimation. A few works have proposed self-supervised training methods [79] or generating pseudo ground truth depth maps [60] to overcome this limitation.

In contrast, *optimization-based methods leveraging perspective foundation models* [47, 42] offer an alternative approach that does not require labeled datasets. Recently, monocular depth estimation methods based on foundation models in the perspective image domain [45, 67, 68, 32, 71, 25, 62] have demonstrated robust performance in zero-shot scenarios. 360° images can be transformed into multiple tangential plane images to reduce distortions. This transformation enables respective depth and normal estimation on each tangential plane, which can then be recombined into a single equirectangular projection. Leveraging pre-trained robust foundation models helps address the limitations associated with the lack of labeled 360° datasets. However, a key challenge remains: Monocular depth estimation encounters scale ambiguity when merging independently estimated depth maps from different viewpoints. While several methods [47, 42] for 360° depth estimation propose blending overlapped regions of tangent-plane depth maps from perspective networks, this simple blending overlooks crucial 3D structural information, as shown in Fig. 1. Although the depth maps may exhibit high quality in 2D space, their corresponding 3D point clouds often suffer from 3D structural inconsistencies and performance degradation.

Main Results: We present a novel structure-aware optimization technique for 360° depth estimation using perspective foundation models [71, 25, 11]. Our approach converts an equirectangular projection image into perspective images via cubemap projections, which represent the 360° scene without overlapping regions. Depth and normal estimation are performed using a perspective foundation model. Since monocular depth estimation provides only relative depth information, the depth scales across different faces of the cubemap may be inconsistent. To address this issue, we introduce a novel graph optimization approach that parameterizes the predicted depth maps and surface normals with additional per-face scale parameters. This parameterization ensures depth scale consistency across different perspective images of the cubemap. Our approach demonstrates robust depth estimation performance across diverse datasets, including indoor and outdoor environments such as Matterport3D [5], Stanford2D3D [4], and 360Loc [26]. The main contributions include:

- We propose a robust optimization-based 360° depth estimation method leveraging perspective foundation models to address the scarcity of labeled 360° dataset.
- We introduce a graph optimization formulation that integrates depth and surface normals with additional per-view scale parameters to enforce depth scale consistency. This optimization preserves the 3D structure and demonstrates superior performance in terms of 3D evaluation metrics.
- Most learning-based methods exhibit performance degradation when trained on a dataset that differs from the test scenes. In contrast, our method does not require a training dataset and is unaffected by domain gap.
- We demonstrate the versatility and benefits of our 360° depth estimation approach by applying it to downstream tasks such as feature matching [30] and structure-from-motion [31].

2 Related Work

2.1 Perspective Foundation Models

Foundation models possess remarkable adaptability, allowing them to generalize effectively across different contexts. CLIP [44] has exhibited outstanding versatility across a wide range of vision-language tasks. DINOv2 [39] introduced a powerful visual encoder designed for various downstream 2D vision applications, such as segmentation, object detection, and depth estimation. DUSt3R [63] further explores the applicability of foundation models in 3D vision tasks, leveraging CroCo [64] pre-training to enhance performance. Furthremore, learning-based monocular depth estimation methods have evolved from optimizing training techniques for various model architectures [12, 16, 19, 20, 6–8] to enhancing zero-shot capabilities [45, 46, 11, 67, 68, 32, 71, 25], enabling strong generalization across diverse domains. Notably, Metric3D [71, 25] extends this direction by specifically exploring zero-shot performance for metric depth estimation and surface normal prediction.

2.2 360° Depth Estimation

360° depth estimation suffers from the inherent distortions introduced by the equirectangular projection. OmniDepth [78] leverage Spherical Convolutions [53] to mitigate this projection distortion. Other methods improve performance by combining multiple projection techniques, such as fusing equirectangular and cubemap projections [58, 59, 27] or adopting icosahedral projections [1]. Additionally, dilated convolutions have been explored to enhance feature extraction in distorted panoramic images [77]. Gravity-aligned feature learning has also been proposed to improve 360° depth estimation [54, 43]. Meanwhile, transformer-based approaches [28, 73, 2, 51] have recently gained traction for capturing global panoramic context more effectively. Despite these advancements, supervised learning methods for 360° depth estimation still heavily rely on labeled datasets. Due to the lack of large-scale real-world annotations, these methods are often dependent on synthetic datasets [11, 75, 3], potentially causing performance degradation in real-world scenarios.

To mitigate this issue, some works explore self-supervised learning via view synthesis [57, 79]. Another line of research focuses on leveraging pre-trained perspective monocular depth estimation models for panorama depth estimation. For example, the pre-trained weights of a perspective monocular depth estimation model are fine-tuned on 360° images using distortion-aware convolutional filters [56]. More recently, perspective foundation models have been incorporated, alleviating the dependency on large-scale labeled 360° datasets. 360MonoDepth [47] utilizes foundation models trained on perspective images [45, 46] to predict depth on tangent images, which are later merged using a blending algorithm. Depth Anywhere [60] adopts a different approach by predicting depth for cubemap images using a foundation model [67] and generating pseudo ground truth, which is then refined through affine-invariant loss. However, while these methods incorporate certain distortion-aware techniques, they lack explicit consideration of 3D structural consistency in the estimated depth maps. We propose a novel approach that leverages perspective foundation models to estimate depth while explicitly promoting 3D structure awareness for 360° depth estimation.

3 Our Approach: RPG360

The overall process is illustrated in Fig. 3. In Section 3.1, the Equirectangular Projection (ERP) image captured by a 360° camera is first converted into distortion-free perspective (PER) images for individual monocular depth estimation. We adopt a cubemap projection to represent PER images due to its computational efficiency, as every pixel is treated as a parameter and the cubemap representation avoids overlapping regions. The estimated depth maps are then merged back into the ERP domain using the refinement method described in Section 3.2.

3.1 Transformation of Coordinate Systems

An ERP image is a 2D plane representation that depicts normalized rays on a spherical surface in a flattened form. The coordinate transformation can be formulated as $\boldsymbol{\xi} = \boldsymbol{\pi}(\boldsymbol{S})$, where $\boldsymbol{\xi} = (\theta, \phi)$ denotes the spherical coordinates with $\theta \in [-\pi, \pi]$, $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. The corresponding 3D Cartesian coordinates are given by the unit vector $\boldsymbol{S} = (S_x, S_y, S_z)$. For

PER images obtained from a six-face cubemap projection, $c \in \{0, ..., 5\}$ indicates the corresponding face of the cubemap. The pixel coordinates of PER images are denoted as p_i,

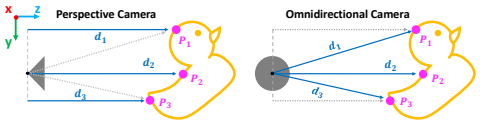

Figure 2: Differences in depth definitions between camera models. (Left) In a perspective camera, depth is defined as the distance along the z-axis from the camera center. (Right) In an omnidirectional camera, depth is measured as the Euclidean distance from the camera center to the 3D point.

$$\begin{cases} S_x = \sin(\theta)\cos(\phi) \\ S_y = \sin(\phi) \\ S_z = \cos(\theta)\cos(\phi), \end{cases} \qquad \widetilde{\boldsymbol{p}}_c = \boldsymbol{K} \cdot \boldsymbol{R}_c^T \cdot \boldsymbol{S}. \quad (1)$$

where $\widetilde{\boldsymbol{p}}_c$ is the homogeneous coordinate of $\boldsymbol{p}_c$. The rotational extrinsic parameter for each face i is represented as $\boldsymbol{R}_c$, while $\boldsymbol{K} \in \mathbb{R}^{3\times3}$ denotes the intrinsic camera matrix, which sets the focal length and the principal point at half of the image size. Using this coordinate transformation, we can convert the depth maps between PER and ERP space. Notably, the definition of depth differs between perspective cameras and omnidirectional cameras. As illustrated in Fig.2, depth in PER cameras is measured as the distance along the z-direction, whereas in 360° cameras, it represents the radial distance from the sensor. Thus, a scaling factor ρ should be applied during the warping process to compensate for the variation in depth definitions,

$$\boldsymbol{D}^{ERP}\langle\boldsymbol{\xi}\rangle = \rho \boldsymbol{D}_c^{PER}\langle\boldsymbol{p}_c\rangle, \qquad \rho = \|\boldsymbol{K}^{-1} \cdot \widetilde{\boldsymbol{p}}_c\|. \quad (2)$$

where $\boldsymbol{D}_i^{PER}$ is the predicted depth map from the pre-trained PER depth estimation model [25] and $\langle\cdot\rangle$ represents the nearest neighbor interpolation. We can obtain 3D points $\boldsymbol{P}$ from the ERP depth map $\boldsymbol{D}^{ERP}$,

$$\boldsymbol{P} = \boldsymbol{D}^{ERP}\boldsymbol{S}. \quad (3)$$

3.2 Graph-based Optimization for Scale Alignment

Scale and Shift Alignment Monocular depth estimation inherently suffers from scale ambiguity, leading to inconsistent depth scales across different PER images. To address this inconsistency, previous studies have proposed methods such as estimating scale and shift between frames [45, 70, 23]. This can be formulated as a linear equation,

$$\boldsymbol{D}^* = \lambda\, \boldsymbol{D} \;+\; \tau. \quad (4)$$

However, ensuring consistent depth $\boldsymbol{D}^*$ across different frames is challenging, as we rely on single scalar values $\lambda \in \mathbb{R}$ and $\tau \in \mathbb{R}$ for all pixels in the depth map $\boldsymbol{D}$.

Graph Optimization for Local Scale Alignment We employ a graph optimization algorithm that parameterizes the predicted depth and normal maps. This optimization refines the depth maps by adjusting the local scale, which is analogous to the shift term, but the adjustment is applied to each pixel rather than relying on a single scalar value. We define this optimization function based on a simple equation that represents a plane defined by $(\boldsymbol{P}_0, \boldsymbol{n}_0)$,

$$\boldsymbol{n_0} \cdot (\boldsymbol{P} - \boldsymbol{P}_0) = 0, \quad (5)$$

where $\boldsymbol{P}_0$ is a 3D point and $\boldsymbol{n}_0$ is its corresponding normal vector of the surface. $\boldsymbol{P}$ indicates any point lying on the same plane. Inspired by [50], we define the 3D points corresponding to each pixel in the predicted depth maps as nodes and assign edge weights based on the likelihood that two adjacent points belong to the same plane,

$$L_p = \sum_i \sum_{j\sim i} w_{ij} \|\mathbf{n}_i \cdot (\mathbf{P}_j - \mathbf{P}_i)\|_2 \;+\; \alpha \sum_i \sum_{j\sim i} w_{ij} \|\mathbf{n}_j - \mathbf{n}_i\|_2, \quad (6)$$

where $j \sim i$ denotes the neighbor pixels of i in the graph, and w_{ij} represents the edge weight between pixel i and its adjacent pixel j in the ERP space. We set α as 0.5. In Eq. 6, the first term ensures two points are positioned in the same plane, while the second term ensures that neighboring points maintain a consistent plane based on the likelihood w_{ij} between adjacent pixels. Based on the assumption that areas with the same texture correspond to the same object surface [34, 14, 49, 50], we define the edge weights using local patches and the spatial distance between pixels,

$$w_{ij} = \exp\left(-\frac{\|\mathbf{Q}_i - \mathbf{Q}_j\|_F^2}{2\sigma_{int}^2}\right) \exp\left(-\frac{\|i-j\|_2^2}{2\sigma_{spa}^2}\right), \quad \text{where} \quad j_x \equiv j_x \bmod W, \;\; j_y = j_y, \quad (7)$$

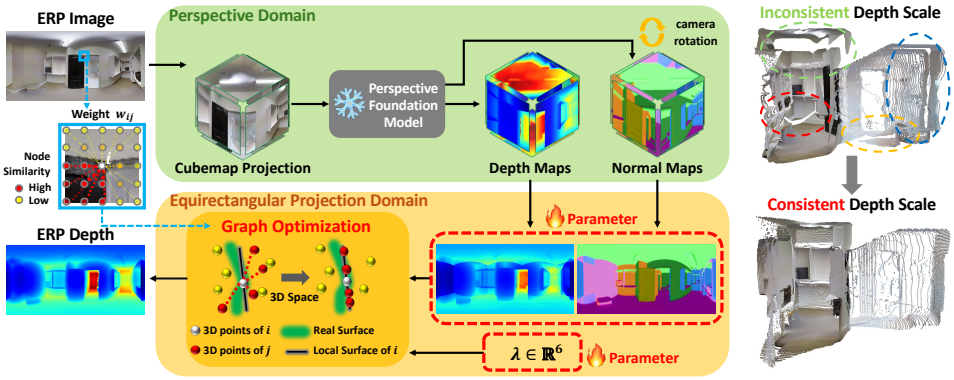

Figure 3: Overview of our depth estimation method. 1) The input equirectangular (ERP) image is projected onto a six-face cubemap, where each non-overlapping face is independently processed by a perspective foundation model to generate depth and normal maps. The normal maps are transformed into a common coordinate system and merged into a single ERP normal map. However, due to the inherent limitations of monocular depth estimation, the merged depth maps exhibit depth scale inconsistencies across faces. 2) To resolve this issue, the merged depth and normal maps are parameterized, incorporating an additional scale parameter λ. Using these parameters, we perform graph optimization under a local planar approximation to obtain a scale-aligned ERP depth map. For the construction of graph, we use the intensity values and the distances between a pixel i (white node) and its adjacent pixels j (red nodes) to define the edge weights w_{ij}. Based on these weights, 3D points are aligned on the surface using both depth and normal information.

where $\mathbf{Q}_i$ represents a patch centered at pixel i of the image, $\|\cdot\|_F$ denotes the Frobenius norm, and σ_{int} and σ_{spa} are set to 0.07 and 3.0, respectively. Since an ERP image has its left and right extremities aligned, the x-coordinate j_x of pixel j should be updated using the modulo operation with the ERP image width W to ensure periodic boundary conditions.

Per-face Parameter for Global Scale Alignment Simply using Eq. 6 may result in over-smoothed depth propagation in neighboring points, as shown in Fig. 6. Therefore, we exploit regularization terms for depth and normal maps to preserve the high fidelity of the input geometry. Here, we incorporate the per-face scale parameter $\lambda_c \in \mathbb{R}^6$ to achieve global scale alignment for the depth of each cubemap face. This additional parameter λ_c functions similarly to the scale term in Eq. 4. The per-face scale parameter λ_c is expanded to match the size of each cubemap face, and then transformed and integrated into the ERP space as a scale map $\boldsymbol{\lambda} \in \mathbb{R}^{H \times W}$.

$$L_d = \sum_i |\boldsymbol{D}_i - \boldsymbol{\lambda}\overline{\boldsymbol{D}}_i|\boldsymbol{m}_i, \quad L_n = \sum_i |\boldsymbol{n}_i - \overline{\boldsymbol{n}}_i|\boldsymbol{m}_i, \tag{8}$$

where $\overline{\boldsymbol{D}}$ and $\overline{\boldsymbol{n}}$ denote input ERP depth and normal maps. The confidence masks $\boldsymbol{m}$ are determined based on the cosine similarity between the input normal and the normal computed from the input depth map using the 8-neighbor convention [69, 29]. The total loss function is formulated as a weighted sum of the proposed loss terms,

$$L_{total} = \eta_p\, L_p + \eta_d\, L_d + \eta_n\, L_n, \tag{9}$$

where η_p, η_d, η_n are weights selected through grid search.

4 Experiments

4.1 Experiments Settings

Datasets and Evaluation Metrics We conduct extensive experiments on benchmark datasets — **Matterport3D** [5], **Stanford2D3D** [4], and **360Loc** [26] — using images of a resolution 1024×512. For Matterport3D and Stanford2D3D, we adopt the official train and test splits. We additionally evaluate on Matterport3D-2K (2048×1024) for high-resolution benchmarks, following [47]. 360Loc consists of four scenes, including both indoor and outdoor environments, each providing panoramic sequences of mapping and query images. We employ all mapping sequences from 360Loc, which include ground truth poses and depth maps, to evaluate the zero-shot performance of depth estimation. As monocular depth estimation increasingly emphasizes 3D

Table 1: Quantitative comparison on the Matterport3D (**M**) and Stanford2D3D (**S**) test set. $\mathbf{M}^+$ indicates the combination of **M** and Structured3D [75]. We evaluate 360° depth estimation performance under three different settings: (a) when the training and test scenes are identical, (b) when the training and test scenes differ, and (c) in a training-free setup. The best methods for each category are shown in bold faces. Most learning-based methods exhibit performance degradation when trained on a dataset that differs from the test scenes. In contrast, the optimization-based methods using perspective foundation models (🤖) do not require a training dataset, making them unaffected by domain gap. Under the training-free setting, RPG360 improves the Chamfer distance by 46.7% on Matterport3D (**M**) and 64.2% on Stanford2D3D (**S**), compared to 360MonoDepth.

	Method	Dataset Train → Test	3D metrics Chamfer ↓	F-Score ↑	IoU ↑
(a)	SliceNet [43]	M → M	1.197	21.247	13.104
	UniFuse [27]	M → M	0.444	42.721	28.573
	ACDNet [77]	M → M	0.584	44.863	30.785
	BiFuse++ [59]	M → M	0.834	39.383	26.066
	Elite360D [1]	M → M	**0.291**	**67.189**	**54.663**
	Depth Anywhere [60]	M^+ → M	0.420	53.997	39.482
(b)	SliceNet [43]	S → M	0.873	24.433	14.734
	UniFuse [27]	S → M	0.542	33.832	**21.711**
	ACDNet [77]	S → M	0.580	32.277	19.985
	Elite360D [1]	S → M	**0.539**	**34.489**	21.548
(c)	360MonoDepth [47]	🤖 → M	0.632	28.056	16.998
	RPG360 (Ours)	🤖 → M	**0.337**	**58.816**	**44.912**

	Method	Dataset Train → Test	3D metrics Chamfer ↓	F-Score ↑	IoU ↑
(a)	SliceNet [43]	S → S	**0.201**	**70.700**	**57.453**
	UniFuse [27]	S → S	0.331	43.481	28.555
	ACDNet [77]	S → S	0.271	56.832	42.007
	Elite360D [1]	S → S	0.223	66.031	51.764
(b)	SliceNet [43]	M → S	1.205	21.681	12.763
	UniFuse [27]	M → S	0.277	56.150	40.678
	ACDNet [77]	M → S	0.495	43.497	28.586
	BiFuse++ [59]	M → S	0.770	46.650	31.731
	Elite360D [1]	M → S	**0.184**	**73.501**	**60.175**
	Depth Anywhere [60]	M^+ → S	0.498	50.927	35.021
(c)	360MonoDepth [47]	🤖 → S	0.576	25.230	14.672
	RPG360 (Ours)	🤖 → S	**0.206**	**73.917**	**61.880**

structural awareness for practical applications [52, 40], we adopt 3D metrics, such as Chamfer Distance, F-Score and IoU, rather than 2D metrics to better assess improvements in 3D structure and geometry. Further details of the 3D metrics are described in the supplementary materials.

Table 2: Quantitative comparison on 360Loc (**L**). For Depth Anywhere, we train and test on 360Loc using its pseudo labeling ($\mathbf{L_p}$) technique. Among optimization-based methods leveraging perspective foundation models (🤖), RPG360 improves the Chamfer distance by 17.2% compared to 360MonoDepth.

	Method	Dataset Train → Test	3D metrics Chamfer ↓	F-Score ↑	IoU ↑
(a)	SliceNet [43]	M → L	2.941	8.164	4.409
	UniFuse [27]	M → L	2.533	10.373	5.670
	ACDNet [77]	M → L	2.423	10.572	5.728
	BiFuse++ [59]	M → L	2.379	8.682	4.647
	Elite360D [1]	M → L	1.759	14.797	8.565
	Depth Anywhere [60]	M^+ → L	2.226	11.889	6.571
	Depth Anywhere [60]	M^+, L_p → L	**1.457**	**17.145**	**10.112**
(b)	360MonoDepth [47]	🤖 → L	2.274	8.992	4.873
	RPG360 (Ours)	🤖 → L	**1.883**	**19.605**	**11.502**

Comparison Models As also noted in [60], many recent methods [1, 2, 73, 36, 37, 51, 72] have not fully released their pre-trained models or provided their code and implementation details. Some also use different datasets, making direct comparison difficult. Therefore, we evaluate our method on fully accessible benchmarks, comparing it with both learning-based [43, 27, 77, 59, 1, 60] and optimization-based [47] approaches using 3D metrics. In addition, to demonstrate the effectiveness of our approach in relation to recent methods, we also include evaluations on 2D metrics. For each method, We adopt the backbone or pre-trained model that achieved the highest performance as reported in the original paper (e.g., EfficientNet-B5 [55] for Elite360D [1] and BiFuse++ for Depth Anywhere [59]). To mitigate scale ambiguity in monocular depth estimation, we apply median alignment [76] in all evaluations. Further details are provided in the supplementary material.

Implementation Details We use the Adam optimizer [33] to perform gradient descent on a single RTX A5000 GPU. To accelerate convergence, we adopt a multi-scale approach following [50]. The ERP depth map $D^{ERP^l} \in \mathbb{R}^{\lfloor h/2^l \rfloor \times \lfloor w/2^l \rfloor}$ where $l \in \{0, ..., L-1\}$, is downsampled by a factor of 2^l. In this experiment, we set $L = 3$ and use learning rates of $5 \times 10^{l-L}$. Each optimization step is performed for 300, 150, and 30 iterations. The weights of the loss terms η_p, η_d, and η_n are set to 50, 0.5, and 10, respectively.

4.2 Experimental Results

As shown in Table 1, although supervised methods perform well when the training and testing domains are the same, their performance degrades significantly when applied to different domains. Interestingly, Table 1 also reveals the opposite trend: some of the methods, such as Elite360D [1], trained in Matterport3D achieve better performance on the Stanford2D3D test set than those trained on the Stanford2D3D train set. Given that the Matterport3D dataset is much larger than Stanford2D3D, we can infer that the available volume of labeled 360° data in Stanford2D3D might be insufficient for effective learning, depending on network capacity. This demonstrates that the lack of labeled 360° datasets poses a significant challenge for supervised learning approaches.

As a potential solution to this problem, the perspective foundation models offer advantages because of their inherent robustness across diverse scenes, and they leverage knowledge from large-scale datasets. 360MonoDepth [47], which employs the perspective foundation model with a blending algorithm,

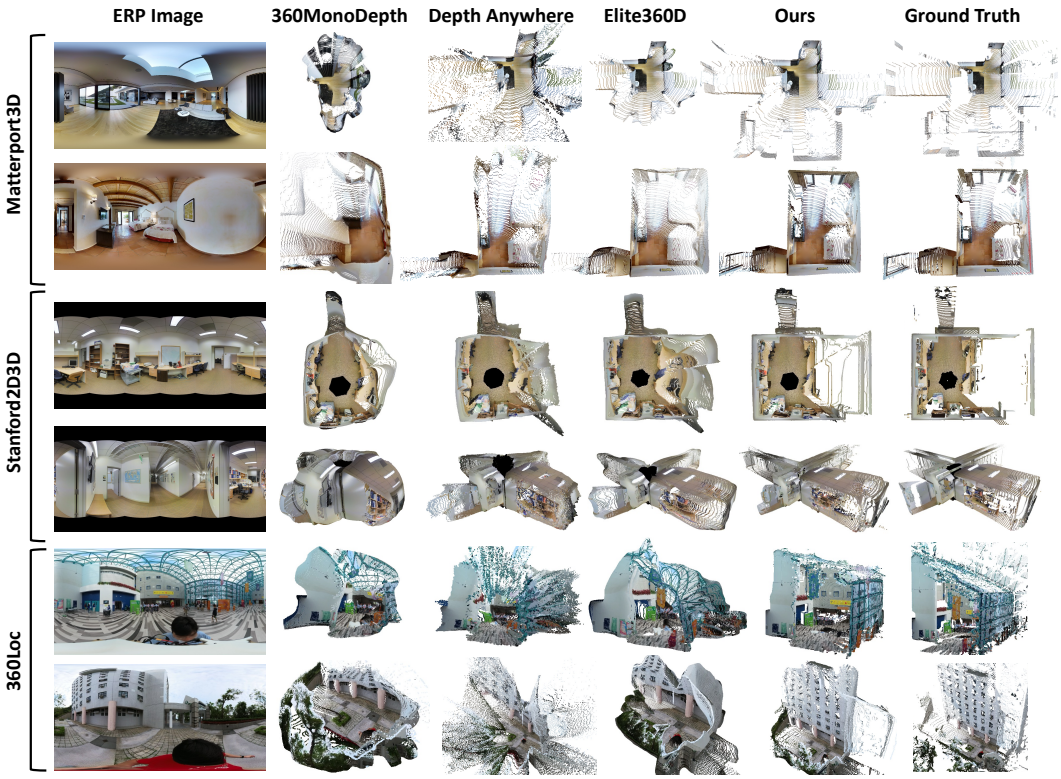

Figure 4: Qualitative comparison of 3D point clouds reconstructed from depth maps across different datasets. Depth Anywhere and Elite360D are learning-based methods, while 360MonoDepth and Ours are optimization-based methods with perspective foundation models. Our RPG360 demonstrates superior structural accuracy, most closely matching the ground truth in 3D space.

achieves high-quality depth estimation in 2D-based metrics [47]. However, its performance in 3D metrics is suboptimal, as it ignores 3D structural information during optimization. Depth Anywhere [60] addresses the lack of labeled datasets by generating a pseudo ground truth through a perspective foundation model. While it demonstrates competitive performance compared to other supervised learning methods, its structural information appears distorted when visualized in 3D point clouds, as shown in Fig. 4. By incorporating our graph-based optimization with the additional per-face parameter, we demonstrate that our method achieves comparable or superior performance across diverse datasets. Furthermore, Fig. 4 shows our method produces point clouds with significantly fewer artifacts in 3D space, further validating its effectiveness.

Table 2 presents zero-shot quantitative results on 360Loc, which consists of both indoor and outdoor scenes. For a fair comparison, we trained Depth Anywhere [60] using its pseudo-labeling method on 360Loc and evaluated its performance. Elite360D [1] performs well in indoor scenes like its training domain but exhibits performance degradation in zero-shot tests. Our method achieves better performance in terms of F-Score and IoU, demonstrating its robustness in diverse environments. We conduct 2D metric evaluations on the Matterport3D dataset as shown in Table 3. Learning-based methods generally demonstrate stronger performance on 2D metrics than optimization-based methods. We also evaluate optimization-based methods using high-resolution 2K images, following [47, 42]. To ensure a fair comparison, we implement 360MonoDepth [47] with Metric3D v2 [25]. However, since 360MonoDepth is originally designed to output inverse depth (as in MiDaS v3 [46]), using a model that predicts standard depth results in degraded blending performance. In contrast, our method requires depth predictions. Therefore, we only evaluate methods that directly produce depth outputs [25, 11], in order to avoid additional normalization issues when converting inverse depth to depth. Since Peng and Zhang [42] leverages both a perspective and a 360° network, it outperforms 360MonoDepth, which relies solely on a perspective network. Although our method relies exclusively on a perspective model, it still achieves superior performance, benefiting from both the robustness of foundation models and the structural awareness introduced by graph optimization.

Table 3: Quantitative comparison using 2D metrics on the (a, b) Matterport3D (**M**, 1024 × 512) and (c, d) Matterport3D-2K (**M-2K**, 2048 × 1024) test sets. In the 2K high-resolution setting, (c) learning-based methods (360 Train.) exhibit performance degradation, whereas (d) optimization-based methods (Opti.), including ours, demonstrate stronger performance.

	Method	Backbone or Pretrained Model	360 Train.	Opti.	Dataset Train → Test	Lower is better Abs Rel	RMSE	Higher is better $\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
(a)	SliceNet [43]	ResNet50 [24]	✓	✗	M → M	0.176	0.613	0.872	0.948	0.972
	UniFuse [27]	ResNet34 [24]	✓	✗	M → M	0.106	0.494	0.890	0.962	0.983
	EGFormer [73]	Transformer	✓	✗	M → M	0.147	0.603	0.816	0.939	0.974
	ACDNet [77]	ResNet50 [24]	✓	✗	M → M	0.101	0.463	0.900	0.968	0.988
	BiFuse++ [59]	ResNet34 [24]	✓	✗	M → M	0.112	0.485	0.881	0.966	0.987
	HRDFuse [2]	ResNet34 [24]	✓	✗	M → M	0.117	0.503	0.867	0.962	0.985
	Elite360D [1]	EfficientNet-B5 [55]	✓	✗	M → M	0.105	**0.452**	0.899	0.971	**0.991**
	Depth Anywhere [60]	BiFuse++ [59], Depth Anything [67]	✓	✗	M^+ → M	**0.085**	-	**0.917**	**0.976**	**0.991**
(b)	360MonoDepth [47]	MiDaS v2 [46]	✗	✓	🤖 → M	0.264	0.916	0.612	0.854	0.941
	RPG360 (Ours)	Omnidata v2 [11]	✗	✓	🤖 → M	0.215	0.672	0.796	0.935	0.973
	RPG360 (Ours)	Metric3D v2 [25]	✗	✓	🤖 → M	**0.203**	**0.667**	**0.859**	**0.953**	**0.977**
(c)	Elite360D [1]	EfficientNet-B5 [55]	✓	✗	M → M-2K	0.309	0.979	0.538	0.824	0.930
	Depth Anywhere [60]	BiFuse++ [59], Depth Anything [67]	✓	✗	M^+ → M-2K	0.167	0.789	0.815	0.947	0.978
	Peng and Zhang [42]	UniFuse [27], LeRes [70]	✓	✓	M → M-2K	**0.115**	**0.611**	**0.871**	**0.954**	**0.982**
(d)	360MonoDepth [47]	MiDaS v3 [46]	✗	✓	🤖 → M-2K	0.208	0.791	0.656	0.890	0.961
	360MonoDepth [47]	Metric3D v2 [25]	✗	✓	🤖 → M-2K	0.300	1.144	0.455	0.766	0.897
	RPG360 (Ours)	Omnidata v2 [11]	✗	✓	🤖 → M-2K	0.147	0.545	0.820	0.947	0.980
	RPG360 (Ours)	Metric3D v2 [25]	✗	✓	🤖 → M-2K	**0.107**	**0.464**	**0.899**	**0.969**	**0.987**

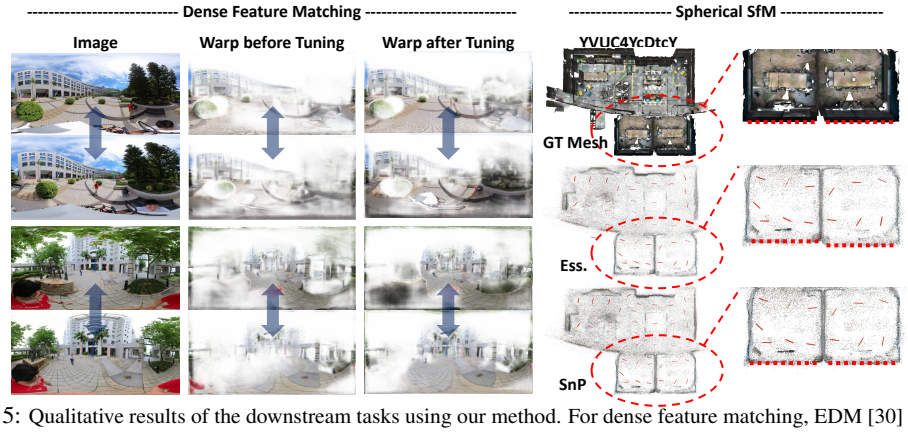

Figure 5: Qualitative results of the downstream tasks using our method. For dense feature matching, EDM [30] estimates more confidential correspondences after tuning with pseudo ground truth generated by RPG360, compared to the initial results. For initial pose estimation in IM360 [31], SnP with RPG360 reconstructs a slightly more accurate 3D structure than two-view geometry-based estimation (Ess.).

4.3 Experimental Analysis

To highlight the significance and utility of robust depth estimation, we apply our method to downstream 360° vision tasks. By leveraging our 360° depth estimation method, we can estimate the relative poses between multiple view frames. To accomplish this, we implement the Spherical-n-Point (SnP) algorithm [21] in conjunction with RANSAC [13].

Dense Feature Matching Since obtaining ground truth correspondences in 360° images is challenging, most existing learning-based feature matching approaches for 360° datasets are primarily restricted to indoor environments [17, 18, 30]. Consequently, when evaluating the feature matching method [30] in outdoor scenes, such as 360Loc, we observe performance degradation. To address this issue, we utilize our robust depth estimation method, which leverages the perspective foundation model. With the aid of SnP and RANSAC, pseudo ground truth correspondences are generated using the relative pose and depth maps from RPG360 in 360Loc query scenes. We then finetune EDM for several iterations and evaluate it on the 360Loc mapping sequences. Table 4 and Figure 5 showcase the advantages of our proposed depth module.

Structure from Motion IM360 [31] demonstrates significant improvements in localization and mapping for large-scale indoor scenes sparsely captured with 360° cameras. During the localization process, relative poses are initialized using two-view geometry estimation based on epipolar geometry and essential matrix decomposition in spherical cameras. Instead of relying on this two-view geometry, we employ SnP [21] with RANSAC [13], utilizing our predicted depth maps before performing spherical incremental triangulation. As shown in Table 6, our approach achieves improvements in AUC errors across most scenes, demonstrating that robust depth estimation can boost the performance of other vision tasks. The qualitative result is presented in Fig. 5.

Table 4: Performance of dense feature matching [30] with and without tuning using RPG360. Our robust depth estimation improves AUC@5° by 5.4% in the hall scene of 360Loc.

Method	Scene	Tuning	AUC ↑ @5°	@10°	@20°
EDM [30]	atrium	✗	37.19	65.01	82.11
EDM [30]	atrium	✓	**38.37**	**65.99**	**82.62**
EDM [30]	concourse	✗	33.13	56.97	75.81
EDM [30]	concourse	✓	**34.53**	**58.52**	**77.26**
EDM [30]	hall	✗	37.83	63.20	80.92
EDM [30]	hall	✓	**39.88**	**64.67**	**81.57**
EDM [30]	piatrium	✗	44.45	69.91	84.86
EDM [30]	piatrium	✓	**46.49**	**70.34**	**85.06**

Table 5: Ablation study on different foundation models and the impact of the per-face parameter λ_c on 3D reconstruction performance. While graph optimization alone provides limited performance improvement, incorporating our proposed per-face parameter λ_c leads to superior results, highlighting its effectiveness in 3D reconstruction quality.

Foundation Model	Graph Opti.	Per-face Param. λ_c	3D metrics Chamfer ↓	F-Score ↑	IoU ↑
Marigold [32]	✓	✓	0.916	23.962	14.262
GeoWizard [15]	✓	✓	0.752	27.923	16.919
Omnidata v2[11]	✓	✓	0.402	50.035	35.968
Metric3D v2[25]	✗	✗	0.380	51.366	36.971
Metric3D v2[25]	✓	✗	0.382	51.605	37.303
Metric3D v2[25]	✓	✓	**0.337**	**58.816**	**44.912**

Table 6: Performance of spherical Structure from Motion (SfM) [31] using Epipolar geometry (Ess.) or Spherical-n-Point (SnP) [21] as the initial pose. RPG360 enables the use of SnP, resulting in a 9.7% improvement in AUC@3° for the YVUC3YcDtcY scene in Matterport3D, demonstrating the effectiveness of our robust depth estimation in enhancing spherical SfM performance.

Method	Scene	Init Pose	AUC ↑ @3°	@5°	@10°
IM360 [31]	2t7WUuJeko7	Ess.	49.16	69.05	84.53
IM360 [31]	2t7WUuJeko7	SnP	**51.17**	**70.48**	**85.24**
IM360 [31]	8194nk5LbLH	Ess.	**34.91**	44.16	66.87
IM360 [31]	8194nk5LbLH	SnP	34.45	**46.56**	**67.25**
IM360 [31]	pLe4wQe7qrG	Ess.	73.95	84.37	92.18
IM360 [31]	pLe4wQe7qrG	SnP	**74.19**	**84.52**	**92.26**
IM360 [31]	RPmz2sHmrrY	Ess.	**53.44**	**71.87**	**85.93**
IM360 [31]	RPmz2sHmrrY	SnP	52.92	71.54	85.77
IM360 [31]	YVUC4YcDtcY	Ess.	72.40	83.40	91.71
IM360 [31]	YVUC4YcDtcY	SnP	**79.44**	**87.67**	**93.83**
IM360 [31]	zsNo4HB9uLZ	Ess.	52.35	71.14	85.49
IM360 [31]	zsNo4HB9uLZ	SnP	**53.50**	**71.84**	**85.86**

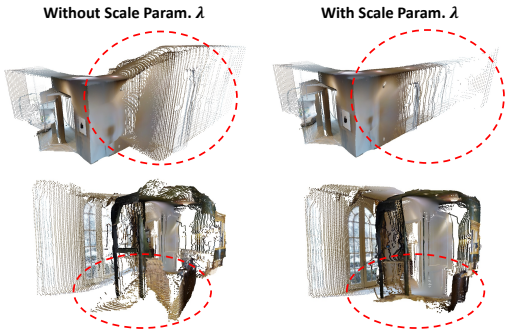

Figure 6: Effect of the per-face scale parameter λ_c. (Left) Without λ_c, Eq. 6 only smooths the gap between 3D points from different face views within the cubemap, leading to noticeable distortions and misaligned structures. (Right) By incorporating λ_c in Eq. 8, depth scale consistency is improved, resulting in enhanced structural integrity.

Ablation Study We analyze the performance variations of our method when using different perspective foundation models, as shown in Table 5. Marigold [32] and GeoWizard [15] utilize Stable Diffusion [48], whereas Omnidata [11] and Metric3D [25] are trained on large-scale datasets using multi-task learning. All these models are capable of predicting both depth and surface normal maps. Although diffusion-based models [32, 15] generate high-fidelity depth maps, their 3D point cloud reconstructions suffer from quality degradation when used for scene reconstruction. Table 5 also demonstrates the effectiveness of the per-face parameter λ_c in graph optimization. Without λ_c, performance remains similar regardless of whether graph optimization is applied. However, incorporating our proposed per-face parameter λ_c leads to noticeable improvements, confirming its contribution to reconstruction quality. The corresponding qualitative improvements are illustrated in Fig. 6. The per-face scale parameter λ_c provides flexibility for 3D points to move collectively within each face, preventing adjacent points from different faces from being simply normalized. This encourages depth consistency while preserving the original structure.

5 Conclusion, Limitations, and Future Work

In this paper, we propose a novel framework for 360° depth estimation. Leveraging foundation models for perspective images, we convert a single ERP image into six-face cubemap images and predict depth and surface normal maps for each face. To ensure depth scale consistency across different faces of the cubemap, we parameterize the predicted depth and normal maps and employ a graph optimization technique with the proposed per-face scale parameter for depth scale alignment. Our optimization-based approach may produce less detailed depth estimations for thin structures compared to training-based methods. However, by leveraging prior knowledge from perspective foundation models, it demonstrates strong robustness in zero-shot settings. Furthermore, as more powerful foundation models emerge, we expect the advantages of our method to become even more pronounced. In the future, we aim to extend the robustness of our approach beyond SfM to applications such as multiview image reconstruction.

Acknowledgements This work was supported in part by ARO Grant W911NF2310352 and Army Cooperative Agreement W911NF2120076.

References

[1] Hao Ai and Lin Wang. Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9926–9935, 2024.

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

[3] Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiros Sterzentsenko, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. Pano3d: A holistic benchmark and a solid baseline for 360deg depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3727–3737, 2021.

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

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

[6] Jaehoon Choi, Dongki Jung, Donghwan Lee, and Changick Kim. Safenet: Self-supervised monocular depth estimation with semantic-aware feature extraction. arXiv preprint arXiv:2010.02893, 2020.

[7] Jaehoon Choi, Dongki Jung, Yonghan Lee, Deokhwa Kim, Dinesh Manocha, and Donghwan Lee. Selfdeco: Self-supervised monocular depth completion in challenging indoor environments. In Proc. of the IEEE International Conference on Robotics and Automation, 2021.

[8] Jaehoon Choi, Dongki Jung, Yonghan Lee, Deokhwa Kim, Dinesh Manocha, and Donghwan Lee. Selftune: Metrically scaled monocular depth estimation through self-supervised learning. In 2022 International Conference on Robotics and Automation (ICRA), pages 6511–6518. IEEE, 2022.

[9] Thiago LT da Silveira, Paulo GL Pinto, Jeffri Murrugarra-Llerena, and Cláudio R Jung. 3d scene geometry estimation from 360 imagery: A survey. ACM Computing Surveys, 55(4):1–39, 2022.

[10] Marek Domański, Olgierd Stankiewicz, Krzysztof Wegner, and Tomasz Grajek. Immersive visual media—mpeg-i: 360 video, virtual navigation and beyond. In 2017 International conference on systems, signals and image processing (IWSSIP), pages 1–9. IEEE, 2017.

[11] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10786–10796, 2021.

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Advances in neural information processing systems, pages 2366–2374, 2014.

[13] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381–395, 1981.

[14] Alessandro Foi and Giacomo Boracchi. Foveated self-similarity in nonlocal image filtering. In Human Vision and Electronic Imaging XVII, volume 8291, pages 296–307. SPIE, 2012.

[15] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In European Conference on Computer Vision, pages 241–258. Springer, 2024.

[16] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In European Conference on Computer Vision, pages 740–756. Springer, 2016.

[17] Christiano Gava, Vishal Mukunda, Tewodros Habtegebrial, Federico Raue, Sebastian Palacio, and Andreas Dengel. Sphereglue: Learning keypoint matching on high resolution spherical images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6134–6144, 2023.

[18] Christiano Gava, Yunmin Cho, Federico Raue, Sebastian Palacio, Alain Pagani, and Andreas Dengel. Spherecraft: A dataset for spherical keypoint detection, matching and camera pose estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 4408–4417, 2024.

[19] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 270–279, 2017.

[20] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE international conference on computer vision, pages 3828–3838, 2019.

[21] Hao Guan and William AP Smith. Structure-from-motion in spherical video using the von mises-fisher distribution. IEEE Transactions on Image Processing, 26(2):711–723, 2016.

[22] Julia Guerrero-Viu, Clara Fernandez-Labrador, Cédric Demonceaux, and Jose J Guerrero. What's in my room? object recognition on indoor panoramic images. In 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 567–573. IEEE, 2020.

[23] S Mahdi H. Miangoleh, Mahesh Reddy, and Yağız Aksoy. Scale-invariant monocular depth estimation via ssi depth. In ACM SIGGRAPH 2024 Conference Papers, pages 1–11, 2024.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[25] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. arXiv preprint arXiv:2404.15506, 2024.

[26] Huajian Huang, Changkun Liu, Yipeng Zhu, Hui Cheng, Tristan Braud, and Sai-Kit Yeung. 360loc: A dataset and benchmark for omnidirectional visual localization with cross-device queries. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22314–22324, 2024.

[27] 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.

[28] Masum Shah Junayed, Arezoo Sadeghzadeh, Md Baharul Islam, Lai-Kuan Wong, and Tarkan Aydın. Himode: A hybrid monocular omnidirectional depth estimation model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5212–5221, 2022.

[29] Dongki Jung, Jaehoon Choi, Yonghan Lee, Deokhwa Kim, Changick Kim, Dinesh Manocha, and Donghwan Lee. Dnd: Dense depth estimation in crowded dynamic indoor scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 12797–12807, 2021.

[30] Dongki Jung, Jaehoon Choi, Yonghan Lee, Somi Jeong, Taejae Lee, Dinesh Manocha, and Suyong Yeon. Edm: Equirectangular projection-oriented dense kernelized feature matching. arXiv preprint arXiv:2502.20685, 2025.

[31] Dongki Jung, Jaehoon Choi, Yonghan Lee, and Dinesh Manocha. Im360: Textured mesh reconstruction for large-scale indoor mapping with 360° cameras, 2025. URL https://arxiv.org/abs/2502.12545.

[32] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9492–9502, 2024.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[34] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. Advances in neural information processing systems, 24, 2011.

[35] Laurent Lescop. 360 vision, from panoramas to vr. In Envisioning architecture: space/time/meaning, volume 1, pages 226–232. Published by the Mackintosh School of Architecture and the School of ..., 2017.

[36] Meng Li, Senbo Wang, Weihao Yuan, Weichao Shen, Zhe Sheng, and Zilong Dong. S2net: Accurate panorama depth estimation on spherical surface. IEEE Robotics and Automation Letters, 8(2):1053–1060, 2023.

[37] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2801–2810, 2022.

[38] Emanuele Menegatti, Takeshi Maeda, and Hiroshi Ishiguro. Image-based memory for robot navigation using properties of omnidirectional images. Robotics and Autonomous Systems, 47(4):251–267, 2004.

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.

[40] Evin Pınar Örnek, Shristi Mudgal, Johanna Wald, Yida Wang, Nassir Navab, and Federico Tombari. From 2d to 3d: Re-thinking benchmarking of monocular depth prediction. arXiv preprint arXiv:2203.08122, 2022.

[41] Gaurav Pandey, James R McBride, and Ryan M Eustice. Ford campus vision and lidar data set. The International Journal of Robotics Research, 30(13):1543–1552, 2011.

[42] Chi-Han Peng and Jiayao Zhang. High-resolution depth estimation for 360deg panoramas through perspective and panoramic depth images registration. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3116–3125, 2023.

[43] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11536–11545, 2021.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

[45] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence, 44(3):1623–1637, 2020.

[46] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF international conference on computer vision, pages 12179–12188, 2021.

[47] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360monodepth: High-resolution 360deg monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3762–3772, 2022.

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.

[49] Mattia Rossi, Mireille El Gheche, and Pascal Frossard. A nonsmooth graph-based approach to light field super-resolution. In 2018 25th IEEE International Conference on Image Processing (ICIP), pages 2590–2594. IEEE, 2018.

[50] Mattia Rossi, Mireille El Gheche, Andreas Kuhn, and Pascal Frossard. Joint graph-based depth refinement and normal estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12154–12163, 2020.

[51] 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.

[52] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter. arXiv preprint arXiv:2208.01489, 2022.

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

[54] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2573–2582, 2021.

[55] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pages 6105–6114. PMLR, 2019.

[56] 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.

[57] Fu-En Wang, Hou-Ning Hu, Hsien-Tzu Cheng, Juan-Ting Lin, Shang-Ta Yang, Meng-Li Shih, Hung-Kuo Chu, and Min Sun. Self-supervised learning of depth and camera motion from 360 videos. In Asian Conference on Computer Vision, pages 53–68. Springer, 2018.

[58] 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.

[59] Fu-En Wang, Yu-Hsuan Yeh, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. IEEE transactions on pattern analysis and machine intelligence, 45(5):5448–5460, 2022.

[60] Ning-Hsu Wang and Yu-Lun Liu. Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. arXiv preprint arXiv:2406.12849, 2024.

[61] 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.

[62] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. arXiv preprint arXiv:2410.19115, 2024.

[63] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20697–20709, 2024.

[64] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 17969–17980, 2023.

[65] Niall Winters, José Gaspar, Gerard Lacey, and José Santos-Victor. Omni-directional vision for robot navigation. In Proceedings IEEE Workshop on Omnidirectional Vision (Cat. No. PR00704), pages 21–28. IEEE, 2000.

[66] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. State-of-the-art in 360 video/image processing: Perception, assessment and compression. IEEE Journal of Selected Topics in Signal Processing, 14(1): 5–26, 2020.

[67] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10371–10381, 2024.

[68] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. arXiv preprint arXiv:2406.09414, 2024.

[69] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[70] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 204–213, 2021.

[71] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9043–9053, 2023.

[72] Ilwi Yun, Hyuk-Jae Lee, and Chae Eun Rhee. Improving 360 monocular depth estimation via non-local dense prediction transformer and joint supervised and self-supervised learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 3224–3233, 2022.

[73] 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.

[74] Fanglue Zhang, Junhong Zhao, Yun Zhang, and Stefanie Zollmann. A survey on 360 images and videos in mixed reality: algorithms and applications. Journal of Computer Science and Technology, 38(3):473–491, 2023.

[75] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, pages 519–535. Springer, 2020.

[76] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1851–1858, 2017.

[77] Chuanqing Zhuang, Zhengda Lu, Yiqun Wang, Jun Xiao, and Ying Wang. Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 3653–3661, 2022.

[78] 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.

[79] 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.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Sec. 1

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 5

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Theory assumptions are not included in this work.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our proposed methods are described in Sec. 3. Details are included in Sec. 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We use the public datasets and cite them in the paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The experiments are done using the fixed random seed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Details are included in Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not use pretrained language models, image generators, or scraped datasets.

    Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Details are included in Appendix.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our proposed method is included in Sec. 3.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification: Crowdsourcing and research with human subjects are not included in this work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This is not related to our work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.