Depth Anywhere: Enhancing 360 Monocular Depth Estimation via Perspective Distillation and Unlabeled Data Augmentation

Ning-Hsu Wang
albert.nhwang@gmail.com

Yu-Lun Liu
Department of Computer Science
National Yang Ming Chiao Tung University
yulunliu@cs.nycu.edu.tw

Abstract

Accurately estimating depth in 360-degree imagery is crucial for virtual reality, autonomous navigation, and immersive media applications. Existing depth estimation methods designed for perspective-view imagery fail when applied to 360-degree images due to different camera projections and distortions, whereas 360-degree methods perform inferior due to the lack of labeled data pairs. We propose a new depth estimation framework that utilizes unlabeled 360-degree data effectively. 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. This method leverages the increasing availability of large datasets. Our approach includes two main stages: offline mask generation for invalid regions and an online semi-supervised joint training regime. We tested our approach on benchmark datasets such as Matterport3D and Stanford2D3D, showing significant improvements in depth estimation accuracy, particularly in zero-shot scenarios. Our proposed training pipeline can enhance any 360 monocular depth estimator and demonstrates effective knowledge transfer across different camera projections and data types. See our project page for results: albert100121.github.io/Depth-Anywhere.

1 Introduction

In recent years, the field of computer vision has seen a surge in research focused on addressing the challenges associated with processing 360-degree images. The widespread use of panoramic imagery across various domains, such as virtual reality, autonomous navigation, and immersive media, has underscored the need for accurate depth estimation techniques tailored specifically for 360-degree images. However, existing depth estimation methods developed for perspective-view images encounter significant difficulties when applied directly to 360-degree data due to differences in camera projection and distortion. While many methods aim to address depth estimation for this camera projection, they often struggle due to the limited availability of labeled datasets.

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. Research efforts on perspective perceptual models have increasingly focused on augmenting the volume of data and developing foundation models that generalize across various types of data. 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 perspective models and large amounts of unlabeled data.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

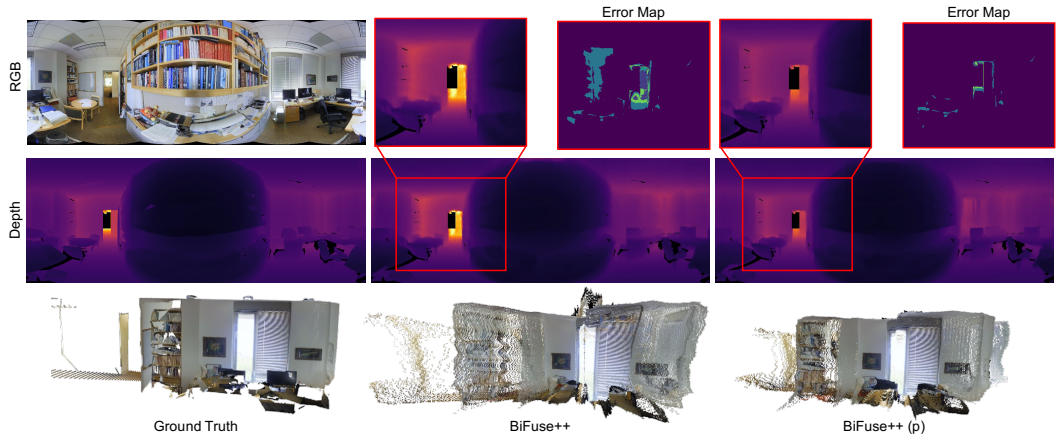

Figure 1: **Our proposed training pipeline improves existing 360 monocular depth estimators.** This figure demonstrated the improvement of our proposed training pipeline tested on the Stanford2D3D [2] dataset in a zero-shot setting.

Our approach consists of two key stages: offline mask generation and online joint training. During the offline stage, we employ a combination of detection and segmentation models to generate masks for invalid regions, such as sky and watermarks in unlabeled data. Subsequently, in the online stage, we adopt a semi-supervised learning strategy, loading half of the batch with labeled data and the other half with pseudo-labeled data. Through joint training with both labeled and pseudo-labeled data, our method achieves robust depth estimation performance on 360-degree imagery.

To validate the effectiveness of our approach, we conduct extensive experiments on benchmark datasets such as Matterport3D and Stanford2D3D. Our method demonstrates significant improvements in depth estimation accuracy, particularly in zero-shot scenarios where models are trained on one dataset and evaluated on another. Furthermore, we demonstrate the efficacy of our training techniques with different SOTA 360-degree depth models and various unlabeled datasets, showcasing the versatility and effectiveness of our approach in addressing the unique challenges posed by 360-degree imagery.

Our contributions can be summarized as follows:

- We propose a novel training technique for 360-degree imagery that harnesses the power of unlabeled data through the distillation of perspective foundation models.
- We introduce an online data augmentation method that effectively bridges knowledge distillation across different camera projections.
- Our proposed training techniques significantly benefit and inspire future research on 360-degree imagery by showcasing the interchangeability of state-of-the-art (SOTA) 360 models, perspective teacher models, and unlabeled datasets. This enables better results even as new SOTA techniques emerge in the future.

2 Related Work

360 monocular depth. Depth estimation for 360-degree images presents unique challenges due to the equirectangular projection and inherent distortion. Various approaches have been explored to address these issues:

- *Directly Apply*: Some methods directly apply monocular depth estimation techniques to 360-degree imagery. OmniDepth [69] leverages spherical geometry and incorporates SphConv [42] to improve depth prediction with distortion. [70, 49] use spherical coordinates to overcome distortion with extra information. [12, 17] leverage other ground truth supervisions to assist on depth estimation. SliceNet [31] and ACDNet [68] propose advanced network architectures tailored for omnidirectional images. EGFormer [64] and HiMODE [18] introduce a transformer-based model that captures global context efficiently, while [45, 44]

focuses on integrating geometric priors into the learning process. [13] proposed to generate large-scale datasets with SfM and MVS, then apply to test-time training.

- *Cube*: Other approaches use cube map projections to mitigate distortion effects. 360-SelfNet [46] is the first work to self-supervised 360 depth estimation leveraging cube-padding [9]. BiFuse [47] and its improved version BiFuse++ [48] are two-branch architectures that utilize cube maps and equirectangular projections. UniFuse [16] combines equirectangular and cube map projections and simplifies the architecture. [3] combines two-branch techniques with transformer network.
- *Tangent Image*: Tangent image projections are also popular. [37, 23, 30] convert equirectangular images into a series of tangent images, which are then processed using conventional depth estimation networks. PanoFormer [40] employs a transformer-based architecture to handle tangent images, while SphereNet [11] and HRDFuse [1] enhance depth prediction by collaboratively learning from multiple projections.

360 other works Beyond depth estimation, 360-degree imagery has been applied to depth completion tasks as follows [26, 8, 32, 57, 58, 15]. Other methods, such as [25] and [43] focus on the projection between camera models, while the former projects pinhole camera model images into a large field of view, whereas the latter transforms convolution kernels.

Unlabeled / Pseudo labeled data. Utilizing unlabeled or pseudo-labeled data has become a significant trend to mitigate the limitations of labeled data scarcity. Techniques like [22, 71, 41, 56] leverage large amounts of unlabeled data to improve model performance through semi-supervised learning. In the context of 360-degree depth estimation, our approach generates pseudo labels from pre-trained perspective models, which are then used to train 360-degree depth models effectively.

Zero-shot methods. Zero-shot learning methods aim to generalize to new domains without additional training data. [7, 54] target this directly with increasing training data., MiDaS [35, 5, 34] and Depth Anything [59] are notable for their robust monocular depth estimation across diverse datasets leveraging affine-invariant loss. [61] takes a step further to investigate zero-shot on metric depth. Marigold [19] leverages diffusion models with image conditioning and up-to-scale relative depth denoising to generate detailed depth maps. ZoeDepth [4] further these advancements by incorporating scale awareness and domain adaptation. [14, 50] leverage camera model information to adapt cross-domain depth estimation.

Foundation models. Foundation models have revolutionized various fields in AI, including natural language processing and image-text alignment. In computer vision, models like CLIP [33] demonstrate exceptional generalization capabilities. [28] proposed a foundation visual encoder for downstream tasks such as segmentation, detection, depth estimation, etc. [20] proposed a model that can cut out masks for any objects. Our work leverages a pre-trained perspective depth estimation foundation model [59] as a teacher model to generate pseudo labels for 360-degree images, enhancing depth estimation by utilizing the vast knowledge embedded in these foundation models.

3 Methods

In this work, we propose a novel training approach for 360-degree monocular depth estimation models. Our method leverages a perspective depth estimation model as a teacher and generates pseudo labels for unlabeled 360-degree images using a 6-face cube projection. Figure 2 illustrates our training pipeline, incorporating the use of Segment Anything to mask out sky and watermark regions in unlabeled data during the offline stage. Subsequently, we conduct joint training using both labeled and unlabeled data, allocating half of the batch to each. The joint training avoids limiting performance by teacher model. The unlabeled data is supervised using pseudo labels generated by Depth Anything, a state-of-the-art perspective monocular depth foundation model. With the benefit of our teacher model, the 360-degree depth model demonstrates an observable improvement on the zero-shot dataset, as shown in Figure 1.

3.1 Unleashing the Power of Unlabel 360 data

Dataset statistics. 360-degree data has become increasingly available in recent years. However, compared to perspective-view depth datasets, labeling depth ground truths for 360-degree data presents greater challenges. Consequently, the availability of labeled datasets for 360-degree data is considerably smaller than that of perspective datasets.

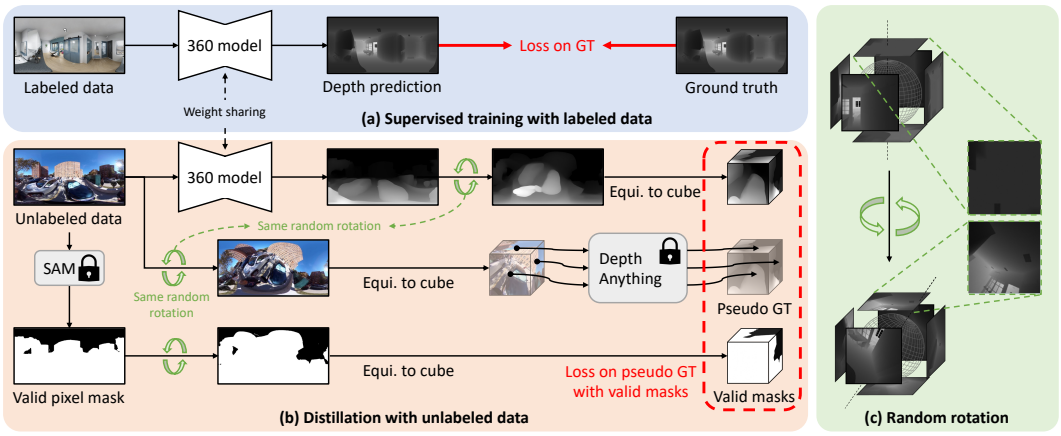

Figure 2: **Training Pipeline.** Our proposed training pipeline involves joint training on both labeled 360 data with ground truth and unlabeled 360 data. (a) For labeled data, we train our 360 depth model with the loss between depth prediction and ground truth. (b) For unlabeled data, we propose to distill knowledge from a pre-trained perspective-view monocular depth estimator. In this paper, we use Depth Anything [59] to generate pseudo ground truth for training. However, more advanced techniques could be applied. These perspective-view monocular depth estimators fail to produce reasonable equirectangular depth as a domain gap exists. Therefore, we distill knowledge by inferring six perspective cube faces and passing them through perspective-view monocular depth estimators. To ensure stable and effective training, we propose generating a valid pixel mask with Segment Anything [20] while calculating loss. (c) Furthermore, we augment random rotation on RGB before passing it into Depth Anything, as well as on predictions from the 360 depth model.

Table 1: **360 monocular depth estimation lacks a large amount of training data.** The number of images in 360-degree monocular depth estimation datasets alongside perspective depth datasets from the Depth Anything methodology.

Perspective				Equirectangular			
Labeled	1.5M	Unlabeled	62 M	Labeled	34K	Unlabeled	344K

Table 1 presents the data quantities available in some of the most popular 360-degree datasets, including Matterport3D [6], Stanford2D3D [2], and Structured3D [65]. Additionally, we list a multi-modal dataset, SpatialAudioGen [29], which consists of unlabeled 360-degree data used in our experiments. Notably, the amount of labeled and unlabeled data used in the perspective foundation model, Depth Anything [59], is significantly larger, with 1.5 million labeled images [24, 52, 10, 60, 55, 51] and 62 million unlabeled images [38, 63, 53, 62, 39, 21, 66, 20], making the amount in 360-degree datasets approximately 170 times smaller.

Data cleaning and valid pixel mask generation Unlabeled data often contains invalid pixels in regions such as the sky and watermark, leading to unstable training or undesired convergence. To address this issue, we applied the GroundingSAM [36] method to mask out the invalid regions. This approach utilizes Grounded DINOv2 [27] to detect problematic regions and applies the Segment Anything [20] model to mask out the invalid pixels by segmenting within the bounding box. While Depth Anything [59] also employs a pre-trained segmentation model, DINOv2, to select sky regions. Brand logos and watermarks frequently appear after fisheye camera stitching. Therefore, additional labels are applied to enhance the robustness of our training process. We also remove all images with less than 20 percent of valid pixels to stablize our training progress.

Perspective foundation models (teacher models). To tackle the challenges posed by limited data and labeling difficulties in 360-degree datasets, we leverage a large amount of unlabeled data alongside state-of-the-art perspective depth foundation models. Due to significant differences in camera projection and distortion, directly applying perspective models to 360-degree data often yields inferior results. Previous works have explored various methods of projection for converting equirectangular to perspective depth, as stated in Section 2. Among these, cube projection and tangent

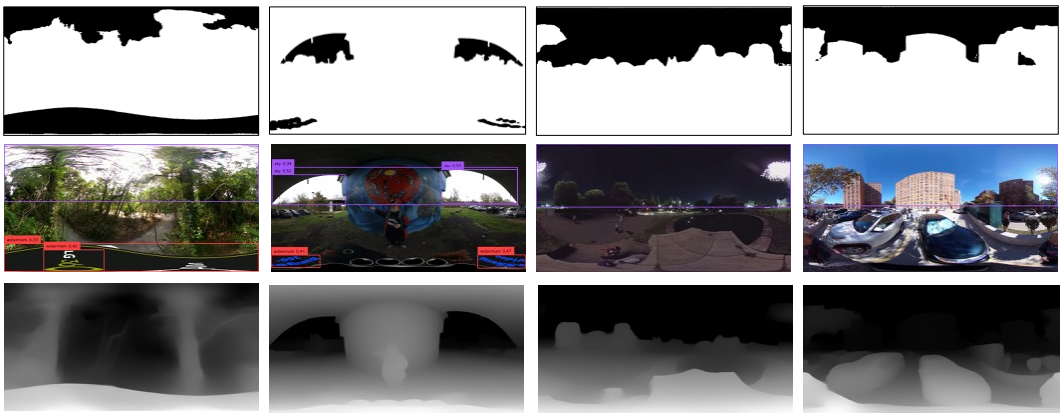

Figure 3: **Valid Pixel Masking.** We used Grounded-Segment-Anything [36] to mask out invalid pixels based on two text prompts: "sky" and "watermark." These regions lack depth sensor ground truth labels in all previous datasets. Unlike Depth Anything [59], which sets sky regions as 0 disparity, we follow ground truth training to ignore these regions during training for two reasons: (1) segmentation may misclassify and set other regions as zero, leading to noisy labeling, and (2) watermarks are post-processing regions that lack geometrical meaning.

projection are the most common techniques. We selected cube projection to ensure a larger field of view for each patch, enabling better observation of relative distances between pixels or objects during the inference of the perspective foundation model and enhancing knowledge distillation. The comparison table can be find in the supplementary material.

In our approach, we apply projection to unlabeled 360-degree data and then run Depth Anything on these projected patches of perspective images to generate pseudo-labels. We explore two directions for pseudo-label supervision: projecting the patch to equirectangular and computing in the 360-degree domain or projecting the 360-degree depth output from the 360 model to patches and computing in the perspective domain. Since training is conducted in an up-to-scale relative depth manner, stitching the patch of perspective images back to equirectangular with an aligned scale will lead to failure in training Figure 4, which is an additional research topic that is worth investigation. We opt to compute the loss in the perspective domain, facilitating faster and easier training without the need for additional alignment optimization.

3.2 Random Rotation Processing

Directly applying Depth Anything on cube-projected unlabeled data does not yield improvements due to ignorance of cross-cube-face relation, leading to cube artifacts (Figure 5). This issue arises from the separate estimation of perspective cube faces, where monocular depth is estimated based on semantic information, lacking a comprehensive understanding of the entire scene. To address this, we propose a random rotation preprocessing step in front of the perspective foundation model.

As depicted in Figure 2, the rotation is applied to equirectangular projection RGB images using a random rotation matrix, followed by cube projection. This results in a more diverse set of cube faces, capturing relative distances between ceilings, walls, windows, and other objects more effectively. With the proposed random rotation technique, knowledge distillation becomes more comprehensive as the point of view is not static. The inference by the perspective foundation model is performed on the fly, with parameters frozen during the training of the 360 model.

In order to perform random rotation, we apply a rotation matrix on the equirectangular coordinates, noted as (θ, ϕ), and rotation matrix as $\mathcal{R}$.

$$(\hat{\theta}, \hat{\phi}) = \mathcal{R} \cdot (\theta, \phi). \tag{1}$$

For equirectangular to cube projection, the field-of-view (FoV) of each cube face is equal to 90 degrees; each face can be considered as a perspective camera whose focal length is $w/2$, and all faces share the same center point in the world coordinate. Since the six cube faces share the same center point, the extrinsic matrix of each camera can be defined by a rotation matrix R_i. p is then the pixel

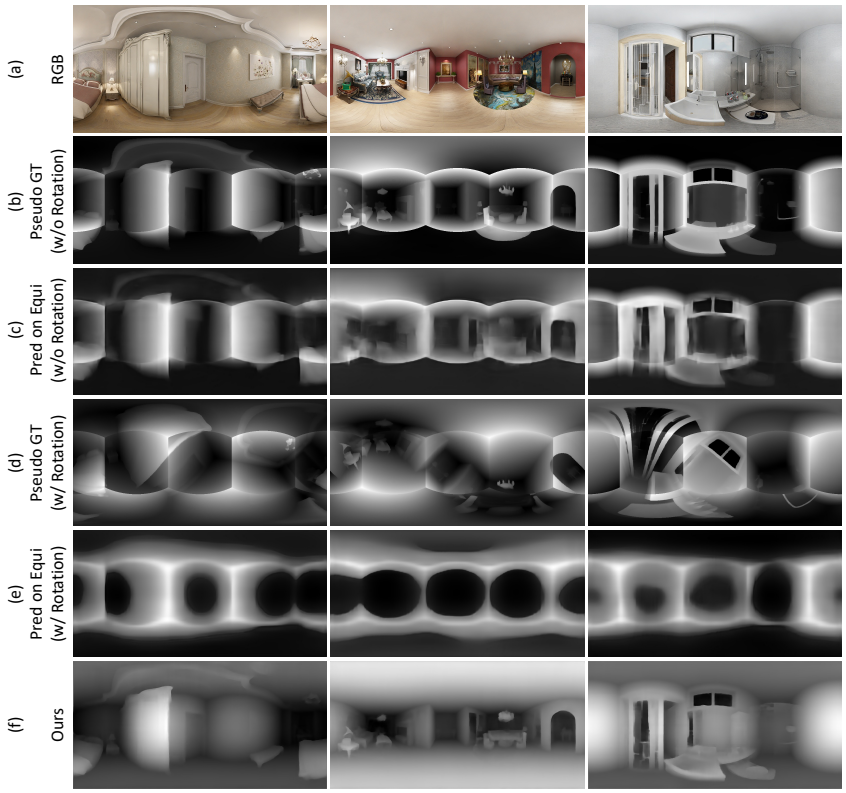

Figure 4: **Qualitative visualization of a model trained directly on pseudo equirectangular data without scale alignment**. We propose calculating the loss with pseudo ground truth on cube faces due to scale misalignment between the six faces during the cube-to-equirectangular projection. We showcase the results of a model trained on pseudo equirectangular data without scale alignment as a simple baseline to demonstrate the importance of calculating loss separately on each of the six faces. The images are presented from top to bottom as follows: (a) RGB images. (b) Pseudo cube ground truth projected directly to equirectangular. (c) Prediction trained with row 2. (d) Pseudo cube ground truth with rotation projected directly to equirectangular. (e) Prediction trained with row 4. (f) Our model's predictions are trained on cube faces separately with rotation.

on the cube face

$$p = K \cdot R_i^T \cdot q, \tag{2}$$

where,

$$q = \begin{bmatrix} q_x \\ q_y \\ q_z \end{bmatrix} = \begin{bmatrix} sin(\theta) \cdot \cos(\phi) \\ \sin(\phi) \\ \cos\theta \cdot \cos\phi \end{bmatrix}, K = \begin{bmatrix} w/2 & 0 & w/2 \\ 0 & w/2 & w/2 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where θ and ϕ are longitude and latitude in equirectangular projection and q is the position in Euclidean space coordinates.

3.3 Loss Function

The training process closely resembles that of MiDaS, Depth Anything, and other cross-dataset methods. Our goal is to provide depth estimation for any 360-degree images. Following previous approaches that trained on multiple datasets, our training objective is to estimate relative depth. The depth values are first transformed into disparity space using the formula $1/d$ and then normalized to the range $[0, 1]$ for each disparity map.

To adapt to cross-dataset training and pseudo ground truths from the foundation model, we employed the affine-invariant loss, consistent with prior cross-dataset methodologies. This loss function disregards absolute scale and shifts for each domain, allowing for effective adaptation across different

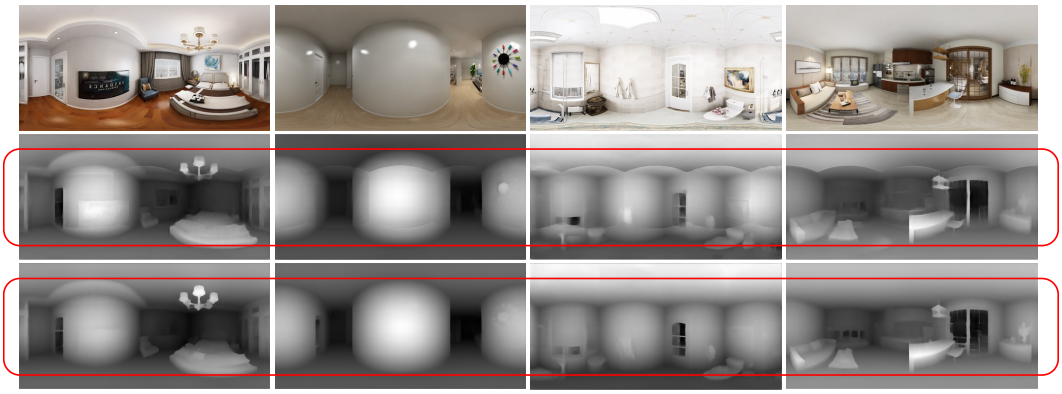

Figure 5: **Cube Artifact.** Shown in the center row of the figure, an undesired cube artifact appears when we apply joint training with pseudo ground truth from Depth Anything [59] directly. This issue arises from independent relative distances within each cube face caused by a static point of view. Ignoring cross-cube relationships results in poor knowledge distillation. To address this, as shown in Figure 2(c), we randomly rotate the RGB image before inputting it into Depth Anything. This enables better distillation of depth information from varying perspectives within the equirectangular image.

datasets and models.

$$\mathcal{L}1 = \frac{1}{HW}\sum_{i=1}^{HW}\rho(d_i^*, d_i), \tag{4}$$

where d_i^* and d_i are the prediction and ground truth, respectively. ρ represents the affine-invariant mean absolute error loss:

$$\rho(d_i^*, d_i) = |\hat{d}_i^* - \hat{d}_i|. \tag{5}$$

Here, $\hat{d}_i$ and $\hat{d}_i^*$ are the scaled and shifted versions of the prediction d_i^* and ground truth d_i:

$$\hat{d}_i = \frac{d_i - t(d)}{s(d)}, \tag{6}$$

where $t(d)$ and $s(d)$ are used to align the prediction and ground truth to have zero translation and unit scale:

$$t(d) = \text{median}(d), \quad s(d) = \frac{1}{HW}\sum_{i=1}^{HW}|d_i - t(d)|. \tag{7}$$

4 Experiments

These notations apply for all tables: **M**: Matterport3D [6], **SF**: Stanford2D3D [2], **ST**: Structured3D [65], **SP**: Spatialaudiogen [29], **-all** indicates using the entire train, validation, and test sets of the specific dataset, and **(p)** denotes using pseudo ground truth generated by Depth Anything [59]. Due to space limits, we provide the experimental setup in the appendix, including implementation details and evaluation metrics.

4.1 Baselines

Recent state-of-the-art methods [1, 64, 47, 48, 16, 31, 40] have emerged. We chose UniFuse and BiFuse++ as our baseline models for experiments, as many of the aforementioned methods did not fully release pre-trained models or provide training code and implementation details. It's worth noting that PanoFormer [40] is not included due to incorrect evaluation code and results. Both selected models are re-implemented with affine-invariant loss on disparity for a fair comparison and to demonstrate improvement. We conduct experiments on the Matterport3D [6] benchmark to demonstrate performance gains within the same dataset/domain, and we perform zero-shot evaluation on the Stanford2D3D [2] test set to demonstrate the generalization capability of our proposed training technique. To further validate its robustness, we evaluate additional baseline models [45, 64] in zero-shot setting, showcasing the effectiveness of our approach for non-dual-projection models.

Table 2: **Matterport3D Benchmark.** The upper section lists 360 methods trained with metric depths in meters using BerHu loss. All numbers are sourced from their respective papers. The lower section includes selected methods retrained with relative depth (disparity) using affine-invariant loss.

Method	Loss	Train	Test	Abs Rel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
BiFuse [47]	BerHu	M	M	-	0.845	0.932	0.963
UniFuse [16]	BerHu	M	M	0.106	0.890	0.962	0.983
SliceNet [31]	BerHu	M	M	-	0.872	0.948	0.972
BiFuse++ [48]	BerHu	M	M	-	0.879	0.952	0.977
HRDFuse [1]	BerHu	M	M	0.097	0.916	0.967	0.984
UniFuse [16]	Affine-Inv	M	M	0.102	0.893	0.970	0.989
UniFuse [16]	Affine-Inv	M, ST-all (p)	M	<u>0.089</u>	0.911	<u>0.975</u>	<u>0.991</u>
BiFuse++ [48]	Affine-Inv	M	M	0.094	<u>0.914</u>	0.974	0.989
BiFuse++ [48]	Affine-Inv	M, ST-all (p)	M	**0.085**	**0.917**	**0.976**	**0.991**

4.2 Benchmarks Evaluation

We conducted our in-domain improvement experiment on the widely used 360-degree depth benchmark, Matterport3D [6], to showcase the results of perspective foundation model distillation on the two selected baseline models, UniFuse [16] and BiFuse++[48]. In Table 2, we list the metric depth evaluation results from state-of-the-art methods on this benchmark. Subsequently, we present the re-trained baseline models using affine-invariant loss on disparity to ensure a fair comparison with their original depth metric training. Finally, we demonstrate the improvement achieved with results trained on the labeled Matterport3D training set and the entire Structured3D dataset with pseudo ground truth.

4.3 Zero-Shot Evaluation

Our goal is to estimate depths for all 360-degree images, making zero-shot performance crucial. Following previous works [47, 16], we adopted their zero-shot comparison setting, where models trained on the entire Matterport3D [6] dataset are tested on the Stanford2D3D [2] test set. In Table 3, the upper section lists methods trained with metric depth ground truth, with numbers sourced from their respective papers. The lower section includes models trained with affine-invariant loss on disparity ground truth. As shown in Figure 6, [16, 48] demonstrate generalization improvements with a lower error on the Stanford2D3D dataset.

Depth Anything [59] and Marigold [19] are state-of-the-art zero-shot depth models trained with perspective depths. As shown in Table 3, due to the domain gap and different camera projections, foundation models trained with perspective depth cannot be directly applied to 360-degree images. We demonstrated the zero-shot improvement on UniFuse [16], BiFuse++ [48] and non-dual-projection methods [45, 64] with models trained on the entire Matterport3D [6] dataset with ground truth and the entire Structured3D [65] or SpatialAudioGen [29] dataset with pseudo ground truth generated using Depth Anything [59].

As Structured3D provides ground truth labels for its dataset, we also evaluate our models on its test set to assess how well they perform with pseudo labels. Table 4 shows the improvements achieved on the Structured3D test set when using models trained with pseudo labels. It's worth noting that even when the 360 model is trained on pseudo labels from SpatialAudioGen, it performs similarly well. This demonstrates the success of our distillation technique and the model's ability to generalize across different datasets.

4.4 Qualtative Results in the Wild

We demonstrated the qualitative results in Figure 8 and Figure 7 360-degree images that were either captured by us or downloaded from the internet[1]. These examples showcase the zero-shot capability of our model when applied to data outside the aforementioned 360-degree datasets.

[1] Stig Nygaard, https://www.flickr.com/photos/stignygaard/49659694937, CC BY 2.0 DEED
Dominic Alves, https://www.flickr.com/photos/dominicspics/28296671029/, CC BY 2.0 DEED
Luca Biada, https://www.flickr.com/photos/pedroscreamerovsky/6873256488/, CC BY 2.0 DEED
Luca Biada, https://www.flickr.com/photos/pedroscreamerovsky/6798474782/, CC BY 2.0 DEED

Table 3: **Zero-shot Evaluation on Stanford2D3D.** We perform zero-shot evaluations with models trained on other datasets. Following the original training settings, we train the 360 models [48, 16, 45, 64] on the entire Matterport3D dataset and then test on Stanford3D's test set.

Method	Loss	train	test	Abs Rel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
BiFuse [47]	BerHu	M-all	SF	0.120	0.862	-	-
UniFuse [16]	BerHu	M-all	SF	0.094	0.913	-	-
BiFuse++ [48]	BerHu	M-all	SF	0.107	0.914	0.975	0.989
Depth Anything [59]	Affine-Inv	Pers.	SF	0.248	0.635	0.899	0.97
Marigold [19]	Affine-Inv	Pers.	SF	0.195	0.692	0.942	0.982
UniFuse [16]	Affine-Inv	M-all	SF	0.090	0.914	0.976	0.990
UniFuse [16]	Affine-Inv	M-all, ST-all (p)	SF	0.086	0.924	0.977	0.990
UniFuse [16]	Affine-Inv	M-all, SP-all (p)	SF	0.090	0.920	0.978	0.990
BiFuse++ [48]	Affine-Inv	M-all	SF	0.090	0.921	0.976	0.990
BiFuse++ [48]	Affine-Inv	M-all, ST-all (p)	SF	**0.082**	**0.931**	**0.979**	0.991
BiFuse++ [48]	Affine-Inv	M-all, SP-all (p)	SF	0.086	0.926	**0.979**	0.991
HoHoNet [45]	Affine-Inv	M-all	SF	0.095	0.906	0.975	0.991
HoHoNet [45]	Affine-Inv	M-all, ST-all (p)	SF	0.088	0.920	**0.979**	**0.992**
EGFormer [64]	Affine-Inv	M-all	SF	0.098	0.906	0.972	0.989
EGFormer [64]	Affine-Inv	M-all, ST-all (p)	SF	0.086	0.923	0.976	0.990

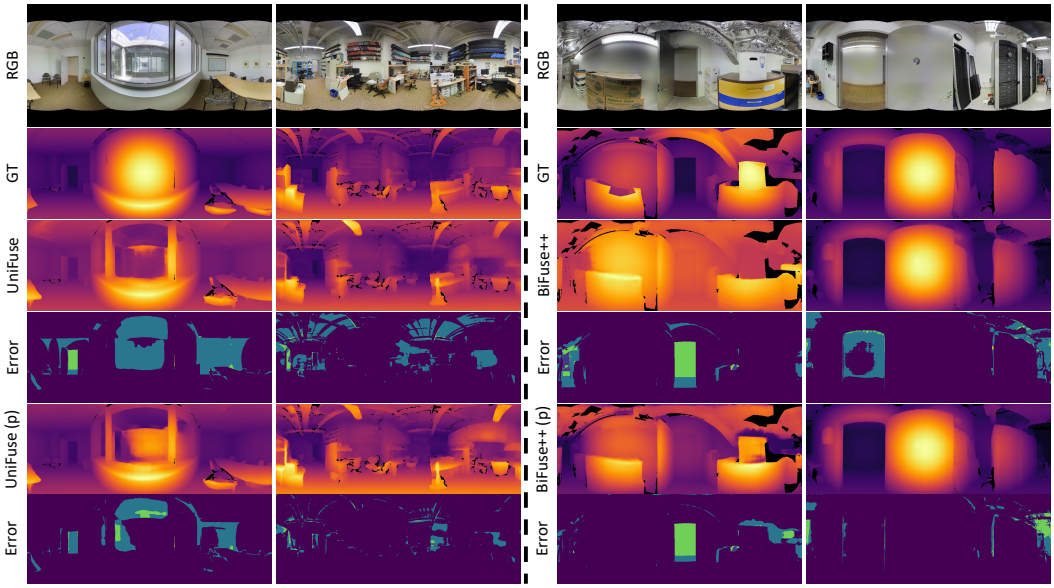

Figure 6: **Zero-shot qualitative with UniFuse [16] (*left*) and BiFuse++ [48] (*right*) tested on Stanford2D3D.**

Table 4: **Structured3D Test Set.** We demonstrate the improvement on the Structured3D test set using pseudo ground truth for training. The lower section shows enhancements with models trained on pseudo ground truth from Matterport3D and SpatialAudioGen, indicating similar improvements. This highlights the successful distillation of Depth Anything.

Method	Loss	train	test	Abs Rel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
UniFuse [16]	Affine-Inv	M-all	ST	0.202	0.759	0.932	0.970
UniFuse [16]	Affine-Inv	M-all, ST-all (p)	ST	0.130	0.887	0.953	0.977
UniFuse [16]	Affine-Inv	M-all, SP-all (p)	ST	0.152	0.864	0.946	0.972

Table 5: **Metric depth fine-tuning.** We fine-tune our model trained with Matterport3D [6] ground truth label and Structured3D [65] pseudo label on relative depth with Stanford2D3D [2]'s training set metric depths with a single epoch.

Method	MAE ↓	Abs Rel ↓	RMSE ↓	RMSElog ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
UniFuse [16]	0.208	0.111	0.369	0.072	0.871	0.966	0.988
UniFuse [16] (Ours)	0.206	0.118	0.351	0.049	0.910	0.971	0.987

Figure 7: **Generalization ability in the wild with depth map visualization.** We showcase zero-shot qualitative results using a combination of images we captured and randomly sourced from the internet to assess the model's generalization ability. For privacy reasons, we have obscured the cameraman in the images.

Figure 8: **Generalization ability in the wild with point cloud visualization.** We showcase zero-shot qualitative results in the point cloud using a combination of images we captured and randomly sourced from the internet to assess the model's generalization ability.

4.5 Fine-Tuned to Metric Depth Estimation

We use our pre-trained model as an initial weight and fine-tune on Stanford2D3D [2] metric depth to demonstrate the effectiveness of our pre-trained relative depth model's ability to adapt to metric depth with a single epoch in Table 5

5 Conclusion

Our proposed method significantly advances 360-degree monocular depth estimation by leveraging perspective models for pseudo-label generation on unlabeled data. The use of cube projection with random rotation and affine-invariant loss ensures robust training and improved depth prediction accuracy while bridging the domain gap between perspective and equirectangular projection. By effectively addressing the challenges of limited labeled data with cross-domain distillation, our approach opens new possibilities for accurate depth estimation in 360 imagery. This work lays the groundwork for future research and applications, offering a promising direction for further advancements in 360-degree depth estimation.

Limitations. Our work faces limitations due to its heavy reliance on the quality of unlabeled data and pseudo labels from perspective foundation models. The results are significantly impacted by data quality (Section 3.1). Without data cleaning, the training process resulted in NaN values. Another limitation is that although with unlabeled data, the scarcity of data still exists compared to other tasks.

References

[1] 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.

[2] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*, February 2017.

[3] Jiayang Bai, Haoyu Qin, Shuichang Lai, Jie Guo, and Yanwen Guo. Glpanodepth: Global-to-local panoramic depth estimation. *IEEE Transactions on Image Processing*, 2024.

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[5] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

[6] 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. *International Conference on 3D Vision (3DV)*, 2017.

[7] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016.

[8] Yichen Chen, Yuqi Pan, Ruyu Liu, Haoyu Zhang, Guodao Zhang, Bo Sun, and Jianhua Zhang. 360orb-slam: A visual slam system for panoramic images with depth completion network. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 717–722. IEEE, 2024.

[9] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. Cube padding for weakly-supervised saliency prediction in 360 {\deg} videos. *arXiv preprint arXiv:1806.01320*, 2018.

[10] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. Diml/cvl rgb-d dataset: 2m rgb-d images of natural indoor and outdoor scenes. *arXiv preprint arXiv:2110.11590*, 2021.

[11] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–533, 2018.

[12] Brandon Yushan Feng, Wangjue Yao, Zheyuan Liu, and Amitabh Varshney. Deep depth estimation on 360 images with a double quaternion loss. In *2020 International Conference on 3D Vision (3DV)*, pages 524–533. IEEE, 2020.

[13] Qi Feng, Hubert PH Shum, and Shigeo Morishima. 360 depth estimation in the wild-the depth360 dataset and the segfuse network. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 664–673. IEEE, 2022.

[14] V. Guizilini, I. Vasiljevic, D. Chen, R. Ambrus, and A. Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9199–9209, Los Alamitos, CA, USA, oct 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.00847. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00847.

[15] Yu-Kai Huang, Tsung-Han Wu, Yueh-Cheng Liu, and Winston H Hsu. Indoor depth completion with boundary consistency and self-attention. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[16] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters*, 2021.

[17] Lei Jin, Yanyu Xu, Jia Zheng, Junfei Zhang, Rui Tang, Shugong Xu, Jingyi Yu, and Shenghua Gao. Geometric structure based and regularized depth estimation from 360 indoor imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2020.

[18] 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.

[19] 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 (CVPR)*, 2024.

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[22] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 2013. URL https://api.semanticscholar.org/CorpusID:18507866.

[23] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Duan Ye, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *2022 Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, June 2022.

[24] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[25] Daniel Lichy, Hang Su, Abhishek Badki, Jan Kautz, and Orazio Gallo. Fova-depth: Field-of-view agnostic depth estimation for cross-dataset generalization. In *2024 International Conference on 3D Vision (3DV)*, pages 1–10, 2024. doi: 10.1109/3DV62453.2024.00056.

[26] Ruyu Liu, Guodao Zhang, Jiangming Wang, and Shuwen Zhao. Cross-modal 360 depth completion and reconstruction for large-scale indoor environment. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):25180–25190, 2022.

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[28] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[29] Timothy Langlois Pedro Morgado, Nuno Vasconcelos and Oliver Wang. Self-supervised generation of spatial audio for 360deg video. In *Neural Information Processing Systems (NIPS)*, 2018.

[30] Chi-Han Peng and Jiayao Zhang. High-resolution depth estimation for 360-degree panoramas through perspective and panoramic depth images registration. *arXiv preprint arXiv:2210.10414*, 2022.

[31] 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 (CVPR)*, pages 11536–11545, June 2021.

[32] Giovanni Pintore, Eva Almansa, Armando Sanchez, Giorgio Vassena, and Enrico Gobbetti. Deep panoramic depth prediction and completion for indoor scenes. *Computational Visual Media*, pages 1–20, 2024.

[33] 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.

[34] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021.

[35] 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), 2022.

[36] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

[37] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360MonoDepth: High-resolution 360 monocular depth estimation. In *CVPR*, 2022.

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[39] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[40] 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.

[41] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

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

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

[44] Cheng Sun, Chi-Wei Hsiao, Ning-Hsu Wang, Min Sun, and Hwann-Tzong Chen. Indoor panorama planar 3d reconstruction via divide and conquer. In *CVPR*, 2021.

[45] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, 2021.

[46] 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*, 2018. URL https://api.semanticscholar.org/CorpusID:53290169.

[47] 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 *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[48] 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, 2023. doi: 10.1109/TPAMI.2022.3203516.

[49] 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.

[50] Ning-Hsu Wang, Ren Wang, Yu-Lun Liu, Yu-Hao Huang, Yu-Lin Chang, Chia-Ping Chen, and Kevin Jou. Bridging unsupervised and supervised depth from focus via all-in-focus supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12621–12631, 2021.

[51] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[52] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020.

[53] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020.

[54] Ke Xian, Chunhua Shen, ZHIGUO CAO, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018. URL https://api.semanticscholar.org/CorpusID:52860134.

[55] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[56] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[57] Zhiqiang Yan, Xiang Li, Kun Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Multi-modal masked pre-training for monocular panoramic depth completion. In *European Conference on Computer Vision*, pages 378–395. Springer, 2022.

[58] Zhiqiang Yan, Xiang Li, Kun Wang, Shuo Chen, Jun Li, and Jian Yang. Distortion and uncertainty aware loss for panoramic depth completion. In *International Conference on Machine Learning*, pages 39099–39109. PMLR, 2023.

[59] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[60] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020.

[61] 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.

[62] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[63] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[64] Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae Eun Rhee. Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6101–6112, October 2023.

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

[66] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[67] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6612–6619, 2017. doi: 10.1109/CVPR.2017.700.

[68] 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.

[69] 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.

[70] Nikolaos Zioulis, Antonis Karakottas, Dimitris Zarpalas, Federic Alvarez, and Petros Daras. Spherical view synthesis for self-supervised 360^o depth estimation. In *International Conference on 3D Vision (3DV)*, September 2019.

[71] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc V. Le. Rethinking pre-training and self-training. *ArXiv*, abs/2006.06882, 2020. URL https://api.semanticscholar.org/CorpusID:219635973.

A Appendix / Supplemental Material

A.1 Experimental Setup

Implementation details. Our work is divided into two stages: (1) offline mask generation and (2) online joint training. (1) In the first stage, we use Grounded-Segment-Anything [36], which combines state-of-the-art detection and segmentation models. We set the `BOX_THRESHOLD` and `TEXT_THRESHOLD` to 0.3 and 0.25, respectively, following the recommendations of the official code. We use "sky" and "watermark" as text prompts. All pixels with these labels are set to False to form our valid mask for the second stage of training. (2) In the second stage, each batch consists of an equal mix of labeled and unlabeled data. We follow the backbone model's official settings for batch size, learning rate, optimizer, augmentation, and other hyperparameters, changing only the loss function to affine-invariant loss. Unlike Depth Anything, which sets invalid sky regions to zero disparity, we ignore these invalid pixels during loss calculation, consistent with ground truth training settings. We average the loss for ground truth and pseudo ground truth during updates. All our experiments are conducted on a single RTX 4090, both offline and online. However, if future 360-degree state-of-the-art methods or perspective foundation models require higher VRAM usage, the computational resource requirements may increase.

Metrics. In line with previous cross-dataset works, all evaluation metrics are presented in percentage terms. The primary metric is Absolute Mean Relative Error (AbsRel), calculated as: $\frac{1}{M}\sum_{i=1}^{M}|a_i - d_i|/d_i$, where M is the total number of pixels, a_i is the predicted depth, and d_i is the ground truth depth. The second metric, δ_j accuracy, measures the proportion of pixels where $max(ai/di, di/ai)$ is within 1.25^j. During evaluations, we follow [16, 47, 48] to ignore areas where ground truth depth values are larger than 10 or equal to 0. Given the ambiguous scale of self-training results, we apply median alignment after converting disparity output to depth before evaluation, as per the method used in [67]:

$$d' = d \cdot \frac{\text{median}(\hat{d})}{\text{median}(d)}, \tag{8}$$

where d is the predicted depth from inverse disparity and $\hat{d}$ is the ground truth depth. This ensures a fair comparison by aligning the median depth values of predictions and ground truths.

A.2 More Qualitative

We demonstrate additional zero-shot qualitative results in Figure 9 and in-the-wild results in Figure 12. In-domain results on the Matterport3D test sets are showcased in Figure 10 and Figure 11.

A.3 Dataset Statistic

As described in Sec.3.1 of the main paper, there is a significant difference in the number of images between the perspective and equirectangular datasets. Detailed statistics of the datasets are listed in Table 6.

A.4 Ground Truth and Pseudo Label Ratio Ablation

Unlike many previous knowledge distillation approaches that use a higher proportion of pseudo labels during model training, we opt for an equal ratio of ground truth to pseudo labels. Through an ablation study exploring the relationship between this data ratio and model performance, we observe a robust improvement starting from $1:1$ in Table 7.

A.5 Perspective Camera Projection Ablation

There are various perspective camera projections for panoramic imagery, with cube and tangent image projections being the most common, both widely used in previous works. In Table 8, we compare these two projections and observe similar improvements when applying our proposed training pipeline, demonstrating the effectiveness of our method. For robust knowledge distillation in relative depth estimation, we select the cube projection due to its wider field of view coverage.

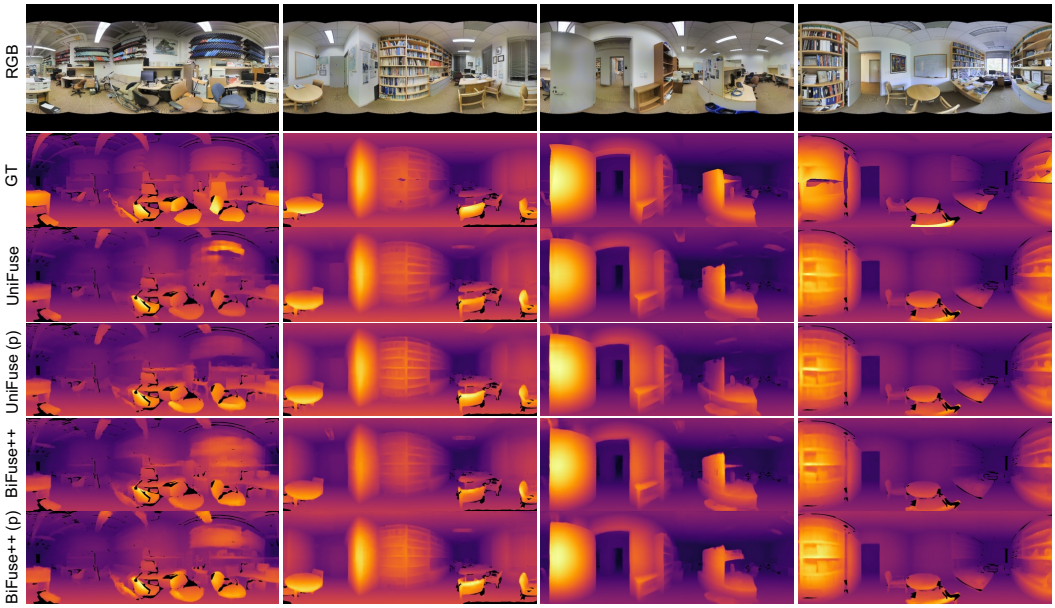

Figure 9: **More qualitative tested on Stanford2D3D with zero-shot setting.**

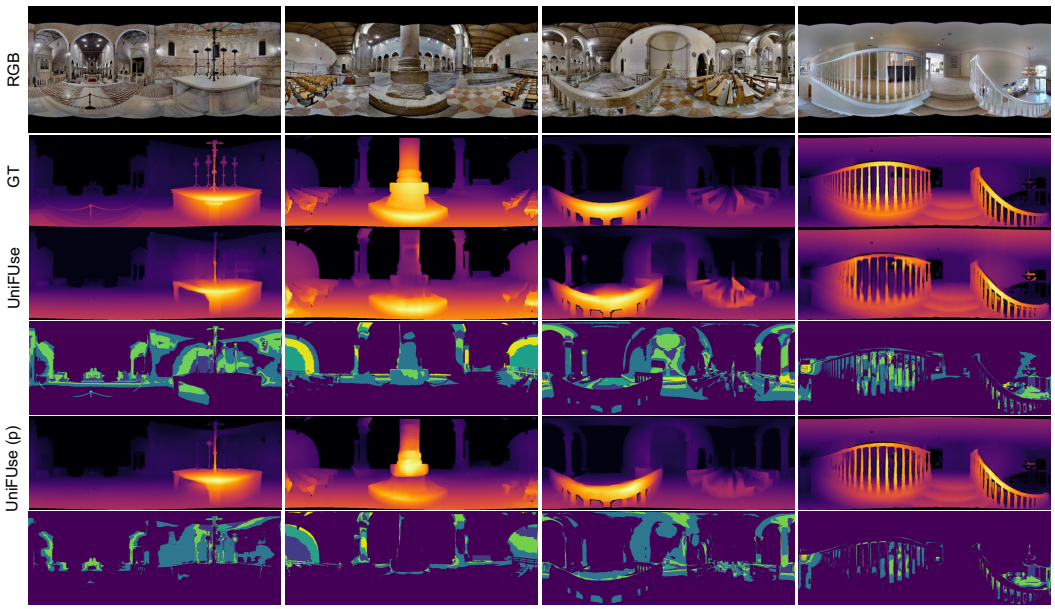

Figure 10: **In-domain qualitative with UniFuse.**

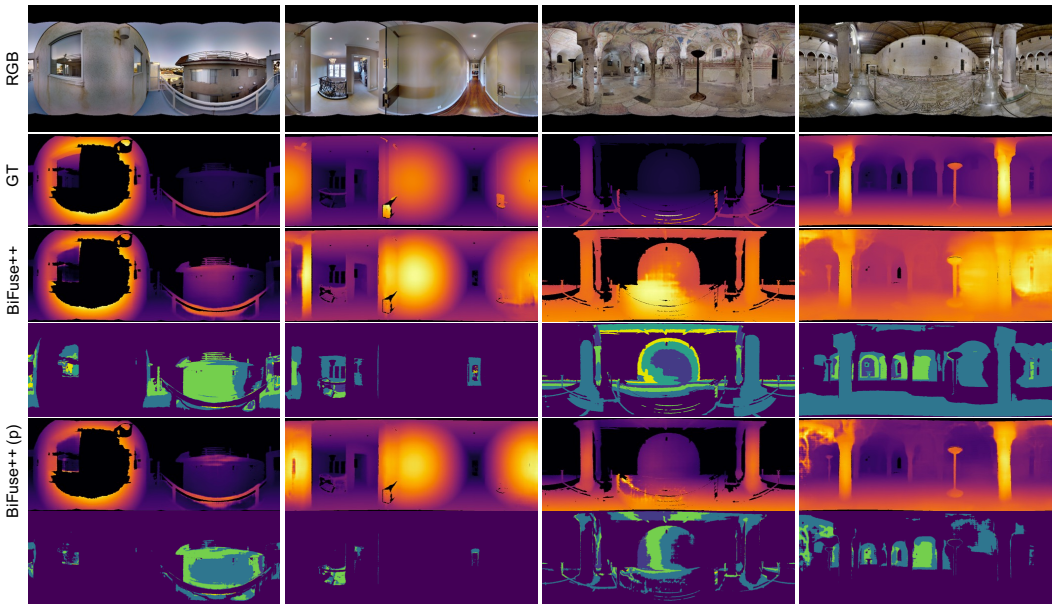

Figure 11: **In-domain qualitative with BiFuse++.**

Table 6: **360 monocular depth estimation lacks a large amount of training data.** This table lists datasets used in 360-degree monocular depth estimation alongside perspective depth datasets from the Depth Anything methodology. The volume of training data for 360-degree imagery (*right column*) is significantly smaller than that for perspective imagery (*left column*) by about 200 times. This highlights the need for using perspective distillation techniques to enhance the limited data available for 360-degree depth estimation. Ground truth (GT) labels are noted where applicable, showing the available resources for training in these domains.

Dataset	Perspective Venue	# of images	GT labels	Dataset	Equirectangular Venue	# of images	GT labels
MegaDepth [24]	CVPR 2018	128K	✓	Stanford2D3D [2]	arXiv 2017	1.4K	✓
TartanAir [52]	IROS 2020	306K	✓	Matterport3D [6]	3DV 2017	10.8K	✓
DIML [10]	arXiv 2021	927K	✓	Structured3D [65]	ECCV 2020	21.8K	✓
BlendedMVS [60]	CVRP 2020	115 K	✓	SpatialAudioGen [29]	NeurIPS 2018	344K	-
HRWSI [55]	CVPR 2020	20K	✓				
IRS [51]	ICME 2021	103K	✓				
ImageNet-21K [38]	IJCV 2015	13.1M	-				
BDD100K [63]	CVPR 2020	8.2M	-				
Google Landmarks [53]	CVPR 2020	4.1M	-				
LSUN [62]	arXiv 2015	9.8M	-				
Objects365 [39]	ICCV 2019	1.7M	-				
Open Images V7 [21]	IJCV 2020	7.8M	-				
Places365 [66]	TPAMI 2017	6.5M	-				
SA-1B [20]	ICCV 2023	11.1M	-				

Table 7: **Ratios of GT and Pseudo label during training.** We conduct additional experiments with varying ratios, which shows our method is robust across different ratios starting from 1:1.

Ratio	train(GT)	train(Pseudo)	test	Abs Rel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
1:1	M-all	ST-all(p)	SF	0.086	0.924	0.977	0.990
1:2	M-all	ST-all(p)	SF	0.087	0.923	0.977	0.990
1:4	M-all	ST-all(p)	SF	0.085	0.923	0.977	0.990

Table 8: **Comparison between Tangent Image and Cube Projection.** We compare two of the most commonly used perspective camera projections for panorama images. As shown in the table, both projections yield similar quantitative results. However, we select the cube projection for knowledge distillation due to its broader field of view coverage.

Projection	Method	train	test	Abs Rel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Cube	UniFuse	M-all+ST-all(p)	SF	**0.086**	**0.924**	0.977	0.990
Tangent	UniFuse	M-all+ST-all(p)	SF	0.087	0.923	**0.978**	**0.991**

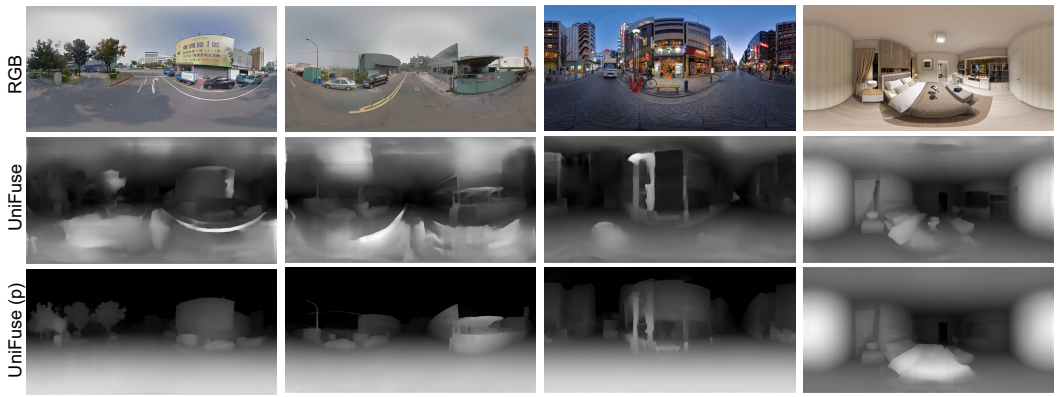

Figure 12: **Additional in-the-wild results.** We compare our proposed joint-training method (Matterport3D (GT) + SpatialAudioGen (Pseudo)) with a model trained only on the Matterport3D dataset, using data randomly downloaded from the internet. This comparison demonstrates the significant improvement of our method along with its generalization ability and effectiveness on real-world data.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Yes, the main claims presented in the abstract and introduction accurately portray the contributions and scope of our paper. As evidenced by Figure 2 and Table 3, our work indeed introduces a training pipeline that demonstrates improvements in the zero-shot testing, making it a significant contribution.

   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: As described in Sec. 5, our work relies heavily on the quality of the data and the accuracy of the pseudo labels. We have also highlighted that disregarding or neglecting data cleaning procedures may result in undesirable training outcomes.

   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: [Yes]

Justification: As indicated in Section 4.3 and Table 3, our train pipeline numerically improves the existing 360 methods. We've also demostrated the effectiveness of pseudo ground truth on Structured3D in Table 4.

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: We have listed all data cleaning preprocessing steps in Sec.3.1, and the implementation details are described in Sec.A.1. This ensures transparency and enables reproducibility of our main experimental results.

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: While the code for the work will be released upon acceptance, we have chosen not to submit it during the reviewing stage. However, as stated in the previous question, we have provided all the necessary information in the paper for reproducibility of the main experimental results.

   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: As specified in Sec. A.1, our training process follows the official implementation of [16, 48], employing an affine-invariant loss and a weighted average between loss from ground truth labels and pseudo labels.

   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: [Yes]

Justification: The error and metric calculations listed in Tables 2, 3, and 4 all adhere to prior works conducted on affine-invariant relative depth on disparity, as described in Sec. A.1.

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: As specified in Sec. A.1, our work was conducted on a single RTX 4090, with the possibility of scaling up using interchangeable state-of-the-art 360 depth methods.

   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: Our work focuses on an AI training pipeline that bridges the gap between different camera models. No extra data are collected, no human-related crowdsourcing is involved, and we address no specific topics that may lead to negative social impact.

   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: [NA]

    Justification: As our work focus on bridge the knowledge distillation between differemce camera projection, it is a relative general research on improving training progress with no negative broader impacts.

    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: As we propose a new training process to bridge the knowledge distillation between camera models, this work does not focus on single model development or dataset collection. Therefore, there are no safeguard issues associated with the release of data or models.

    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: In this work, we've utilized code [48, 16, 59, 20, 36, 27, 45, 64] and dataset [6, 2, 65, 29] from various sources. We've adhered to their respective licenses, cited them appropriately, and incorporated their contributions into our research. For data used in in-the-wild generalization testing, we downloaded data that are free to share and adapt. Their authors, URLs, and licenses are referenced properly in Sec. 4.4.

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 work focuses on improving the training pipeline of previous works [16, 48] with the addition of [59], all of which are publicly accessible and have been fully cited in our work. Additionally, we explain the training procedure in detail in the supplementary materials (Sec. A.1).

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: Our work does not involve any crowdsourcing or human subjects.

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: Our work does not involve any crowdsourcing or human subjects.

    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.