# RePer-360: Releasing Perspective Priors for 360° Depth Estimation via Self-Modulation

Cheng Guan[1,2], Chunyu Lin[1,2], Zhijie Shen[1,2✉], Junsong Zhang[1,2], and Jiyuan Wang[1,2]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China.
[2] Visual Intelligence +X International Cooperation Joint Laboratory of MOE, Beijing 100044, China.
✉ Corresponding author: Zhijie Shen
{24120278,zjshen}@bjtu.edu.cn

**Abstract.** Recent depth foundation models trained on perspective imagery achieve strong performance, yet generalize poorly to 360° images due to the substantial geometric discrepancy between perspective and panoramic domains. Moreover, fully fine-tuning these models typically requires large amounts of panoramic data. To address this issue, we propose RePer-360, a distortion-aware self-modulation framework for monocular panoramic depth estimation that adapts depth foundation models while preserving powerful pretrained perspective priors. Specifically, we design a lightweight geometry-aligned guidance module to derive a modulation signal from two complementary projections (i.e., ERP and CP) and use it to guide the model toward the panoramic domain without overwriting its pretrained perspective knowledge. We further introduce a Self-Conditioned AdaLN-Zero mechanism that produces pixel-wise scaling factors to reduce the feature distribution gap between the perspective and panoramic domains. In addition, a cubemap-domain consistency loss further improves training stability and cross-projection alignment. By shifting the focus from complementary-projection fusion to panoramic domain adaptation under preserved pretrained perspective priors, RePer-360 surpasses standard fine-tuning methods while using only 1% of the training data. Under the same in-domain training setting, it further achieves an approximately 20% improvement in RMSE. The code is available at https://github.com/munimo/RePer360.

**Keywords:** 360° image · Depth Estimation · Domain Adaptation

## 1 Introduction

Unlike traditional perspective images, 360° imagery captures a complete spherical environment in a single shot, enabling holistic scene understanding and broad applications in virtual reality and autonomous systems. Recent depth foundation models, such as Depth Anything Models [35, 36] (DAMs), generalize well on perspective images but degrade markedly on 360° panoramas. We attribute this

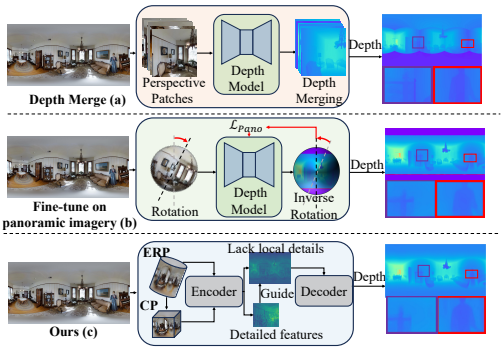


**Fig. 1:** Technical strategies based on pretrained perspective models (Persp): (Top) Patch fusion suffers from artifacts; (Middle) End-to-end fine-tuning requires massive data; (Bottom) Our RePer-360 enables precise detail transfer with minimal data.

drop to a severe prior mismatch: pretrained representations follow perspective-domain statistics that panoramic distortions violate.

To address this challenge, existing solutions can be broadly categorized into two lines of work (see Fig. 1). The first follows a *projection-based inference-and-fusion* paradigm. MoGe-2 [33] adopts this strategy by partitioning a panorama into multiple perspective views, performing inference per view, and fusing the predictions back to the panoramic space. Building upon this pipeline, ST$^2$360D [5] organizes perspective projections into a pseudo-temporal sequence and leverages video foundation models to further improve cross-view consistency. While such pipelines can alleviate local artifacts, they typically treat projected views as approximately independent and do not explicitly model global spherical geometry. Moreover, multi-view inference and the subsequent fusion steps introduce additional computational overhead and inference latency, limiting efficiency.

The second line fine-tunes perspective depth models in the panoramic domain using additional 360° data. For instance, PanDA [6] employs LoRA for parameter-efficient adaptation under a semi-supervised framework. Although this approach can better preserve global geometric consistency, it often relies on large-scale panoramic data. More importantly, without explicitly modeling panoramic distortions, domain adaptation under limited supervision may fail to effectively transfer pretrained perspective priors and can even overwrite them or induce representation drift, ultimately harming generalization.

Inspired by multi-projection strategies, we first explored a complementary-projection feature fusion scheme: we froze the backbone of a depth foundation model, extracted two sets of features from complementary projections, such as Cubemap Projection (CP) and ERP, and applied established fusion strategies to simultaneously handle distortion and transfer perspective priors. However, we observed only marginal gains. We hypothesize that explicit fusion perturbs the pretrained feature statistics, which in turn degrades performance (see Sec. 4.3). Motivated by this finding, we argue that complementary projection features

should *not* be hard-fused into a new representation; instead, the complementary features should be used as a structured guidance signal to enable more stable domain transfer of pretrained priors.

To this end, we propose **RePer-360**, a distortion-aware *self-modulation* framework for panoramic monocular depth estimation. Rather than performing direct feature fusion, RePer-360 aligns cross-domain feature distributions via *normalization-based modulation*, thereby adapting to panoramic distortions while avoiding overwriting or damaging pretrained perspective priors. Concretely, we propose *Self-Conditioned AdaLN-Zero* (SCAdaLN-Zero), a zero-initialized normalization modulator that derives pixel-wise scaling from geometry-aligned guidance and applies it within normalization layers to enable stable adaptation. A lightweight geometry-aligned guidance module extracts modulation cues from complementary projections (CP and ERP), explicitly encoding distortion differences and geometric correspondences. In addition, we design a *cubemap-domain consistency loss* to mitigate distortion-induced imbalance and to enhance cross-projection consistency.

RePer-360 achieves high data efficiency and superior accuracy: using only $\sim$1% of the training data of prior methods (1k vs. 120k image pairs), it surpasses the previous state of the art; under the same training data scale, it improves RMSE by up to 22.4% relatively. Our contributions are summarized as follows:

- We reformulate panoramic adaptation as *distortion-aware, guidance-based domain adaptation*, using complementary projections as guidance for prior-preserving transfer instead of hard fusion.
- We propose RePer-360, a distortion-aware self-modulation framework that performs stable perspective-to-panorama alignment via geometry-aligned guidance and normalization-based modulation.
- Extensive experiments demonstrate that our method is efficient under limited panoramic supervision and yields consistent performance improvements.

## 2 Related work

### 2.1 Monocular 360 Depth Estimation

Although a giant leap has been made in the 3D area [27–31], panoramic distortions significantly degrade the performance of depth estimation models originally designed for perspective images. Early research addressed this challenge through two main approaches: developing fusion frameworks [2, 24, 38] that integrate Cubemap projection (CP) with Equirectangular projection (ERP), and designing specialized attention modules [23, 39, 41] to capture distortion patterns directly in the spherical domain. With the development of large pretrained perspective models such as Depth Anything (DAM), research has shifted toward adapting these powerful priors to panoramic scenarios. Current methodologies follow two technical pathways. The first category divides panoramic images into perspective-aligned patches to alleviate distortions. This strategy is exemplified by MoGe-2 [33] and further extended by ST$^2$360D [5], which arranges projected views

into pseudo-temporal sequences and leverages video-based foundation models for improved consistency. Similar paradigms are adopted by OmniFusion [19] and HRDFuse [1]. Although these methods mitigate local inconsistencies, they process projected views independently without explicitly modeling global spherical geometry. Moreover, their reliance on stitching procedures or multi-frame inference introduces substantial computational overhead and increased latency. The second category focuses on direct fine-tuning of perspective foundation models using large-scale panoramic datasets. PanDA [6] employs LoRA [13] for parameter-efficient semi-supervised fine-tuning of DAM, incorporating a Möbius transformation-based spatial augmentation strategy. Depth Anywhere [32] instead adopts an indirect distillation scheme, generating pseudo-labels from a perspective DAM to supervise panoramic models. However, these approaches share a common limitation: the underlying architectures lack an inherent design for panoramic geometry. Consequently, their performance heavily depends on large-scale 360° supervision, which increases data requirements and may weaken high-fidelity detail preservation. Distortion-aware domain adaptation has also been studied in related panoramic perception tasks such as semantic segmentation [40, 43, 44].

### 2.2 Diffusion Transformer

The Diffusion Transformer [20] (DiT) leverages Adaptive Normalization [21] (AdaLN) for feature-level conditioning, a mechanism also widely adopted in GANs [4, 15] and U-Net based diffusion models [11]. Instead of using fixed parameters, the AdaLN module in DiT dynamically generates channel-level scaling and shifting parameters directly from timestep and class embeddings. A key enhancement introduced in DiT is AdaLN-Zero, which incorporates additional scaling factors that are initialized to zero and applied before residual connections, significantly improving training stability. While these methods condition on externally predefined signals such as timesteps or class embeddings, our approach derives conditioning signals from internally computed, geometry-aligned cross-projection features. This shift replaces generic sequential conditioning with task-specific structural cues tailored to panoramic adaptation.

## 3 Methodology

**Overview.** RePer-360 adapts a pretrained depth foundation model to the panoramic domain while preserving its priors. As illustrated in Fig. 2, our framework follows a guidance–modulation–supervision pipeline. The panorama is first processed by a frozen DAM encoder to extract ERP features, while a cubemap projection (CP) branch provides complementary perspective-aligned features. A Geometry-Aligned Guidance (GAG) module (Sec. 3.1) aligns and aggregates ERP and CP representations to produce geometry-aware conditioning signals. These signals are injected into normalization layers via a Self-Conditioned AdaLN-Zero module (Sec. 3.2), enabling structured distortion-aware adaptation while preserving pretrained representations. To stabilize learning under spherical

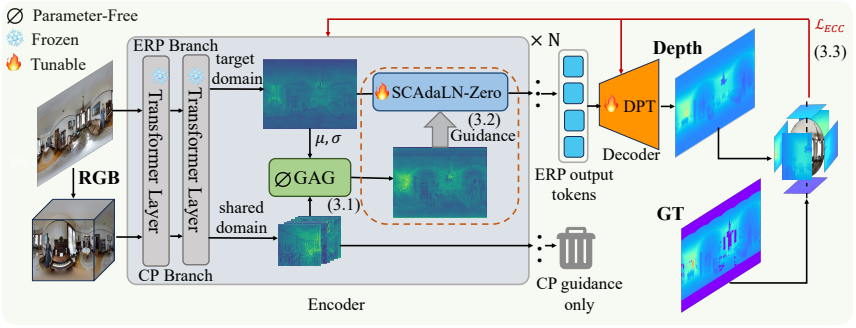


**Fig. 2:** Overview of the proposed framework. Our framework takes a single panorama as input and outputs the corresponding depth map. The network leverages the rich visual priors in the DAM to make it suitable for panoramic depth estimation.

distortion, we further introduce an E2C Consistency Loss (ECCLoss) (Sec. 3.3), which enforces cross-projection geometric consistency in the cubemap domain. Through guidance-driven normalization, RePer-360 reframes panoramic depth estimation as structured prior adaptation rather than explicit feature fusion.

### 3.1 Geometry-Aligned Guidance for Stable Modulation

We introduce a Geometry-Aligned Guidance (GAG) module to generate auxiliary guidance for SCAdaLN-Zero modulation, rather than replacing the backbone feature. Since the backbone is pretrained on perspective images, each cubemap projection (CP) face is closer to the backbone's training distribution than the distorted ERP image. Therefore, CP features are more geometrically consistent with the backbone representation and provide reliable local structural details. However, CP features are split into six faces and suffer from boundary discontinuities, making them unsuitable as a global representation. ERP features, in contrast, preserve the continuous panorama layout and provide stable global context, but they may contain stronger projection distortion. GAG combines these complementary features by first matching the CP feature statistics to the ERP features in the cubemap domain, and then using a learned gate to mix the aligned CP feature with the ERP feature. As shown in Fig. 3, IN denotes instance normalization, and the gate heatmap indicates whether each region relies more on CP or ERP information.

The guidance construction process is implemented as follows. Let $\mathbf{F}_{\text{ERP}} \in \mathbb{R}^{1\times C\times H\times W}$ and $\mathbf{F}_{\text{CP}} \in \mathbb{R}^{6\times C\times H_c\times W_c}$ denote the input features from the ERP and CP branches respectively. Since these two features are defined in different projection domains, we first project $\mathbf{F}_{\text{ERP}}$ into the cubemap domain, yielding $\mathbf{F}_{\text{E2C}}$ for subsequent processing. Next, we perform parameter-free channel-wise statistical alignment between domains via affine normalization. Specifically, we compute the channel-wise mean and standard deviation over spatial dimensions:

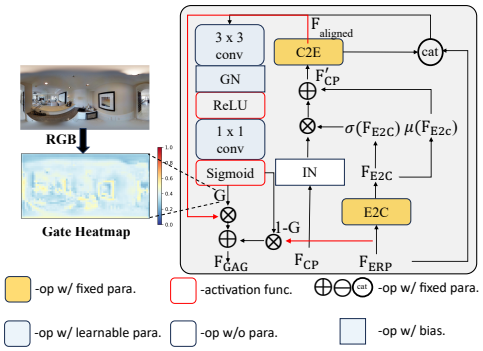


**Fig. 3:** Geometry-Aligned Guidance (GAG). Left: learned gate between aligned CP and ERP features, where red/blue indicates stronger CP/ERP contribution. Right: guidance construction process, where IN denotes instance normalization, E2C ERP-to-cubemap projection, and C2E cubemap-to-ERP projection.

$$\mu_{\mathrm{E2C}} = \overline{\mathbf{F}}^{HW}_{\mathrm{E2C}}, \qquad \sigma_{\mathrm{E2C}} = \sqrt{\overline{\left(\mathbf{F}_{\mathrm{E2C}} - \mu_{\mathrm{E2C}}\right)^2}^{HW} + \epsilon}, \tag{1}$$

$$\mu_{\mathrm{CP}} = \overline{\mathbf{F}}^{HW}_{\mathrm{CP}}, \qquad \sigma_{\mathrm{CP}} = \sqrt{\overline{\left(\mathbf{F}_{\mathrm{CP}} - \mu_{\mathrm{CP}}\right)^2}^{HW} + \epsilon}. \tag{2}$$

Here $\mathbf{F} \in \mathbb{R}^{B\times C\times H\times W}$, and $\overline{(\cdot)}^{HW}$ denotes averaging over spatial locations:

$$\overline{\mathbf{F}}^{HW}_{b,c} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{F}_{b,c,h,w}. \tag{3}$$

This operation can be understood as first normalizing the CP feature using its own channel-wise statistics, and then rescaling it using the statistics of the ERP feature after projection to the cubemap domain. The aligned CP features are obtained by:

$$\mathbf{F'}_{\mathrm{CP}} = \sigma_{\mathrm{E2C}} \cdot \frac{\mathbf{F}_{\mathrm{CP}} - \mu_{\mathrm{CP}}}{\sigma_{\mathrm{CP}}} + \mu_{\mathrm{E2C}}. \tag{4}$$

Here, $\mathbf{F'}_{\mathrm{CP}}$ represents the CP features whose mean ($\mu_{\mathrm{CP}}$) and standard deviation ($\sigma_{\mathrm{CP}}$) have been aligned to match those of the ERP features ($\mu_{\mathrm{E2C}}$ and $\sigma_{\mathrm{E2C}}$). This alignment preserves the local structure of CP features while reducing the distribution gap between CP and ERP features.

To enable subsequent processing in the ERP domain, the calibrated cubemap features $\mathbf{F'}_{\mathrm{CP}}$ are projected back to the original equirectangular coordinate system, producing the geometrically aligned features $\mathbf{F}_{\mathrm{aligned}}$. After this projection, $\mathbf{F}_{\mathrm{aligned}}$ has the same spatial layout as $\mathbf{F}_{\mathrm{ERP}}$, so the two features can be adaptively combined. Building on this alignment, an adaptive gating mechanism performs region-aware feature selection. The gating weights are computed as:

$$\mathbf{G} = \mathrm{Sigmoid}\left(\mathcal{F}_{\mathrm{g}}\left(\left[\mathbf{F}_{\mathrm{aligned}}, \mathbf{F}_{\mathrm{ERP}}\right]\right)\right). \tag{5}$$

where $\mathbf{G} \in [0,1]^{1\times C\times H\times W}$ represents the spatially adaptive weighting, $\mathcal{F}_{\mathrm{g}}$ denotes the gating function, and $[\cdot,\cdot]$ represents the concatenation operation. The gate is predicted from both the aligned CP feature and the ERP feature, allowing the model to choose the more suitable source according to local content.

The final guidance representation is constructed by adaptively aggregating the aligned CP features and the original ERP features through learned gating weights:

$$\mathbf{F}_{\mathrm{GAG}} = \mathbf{G} \odot \mathbf{F}_{\mathrm{aligned}} + (\mathbf{I} - \mathbf{G}) \odot \mathbf{F}_{\mathrm{ERP}}. \tag{6}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{I}$ is an all-ones matrix with the same dimensions as $\mathbf{G}$.

Importantly, $\mathbf{F}_{\mathrm{GAG}}$ is not used as the final backbone representation, but instead parameterizes the subsequent SCAdaLN-Zero modulation. By preserving CP-derived local structures while incorporating ERP-based global context, it provides geometry-aware guidance for structure-aware adaptation.

### 3.2 Self-Conditioned AdaLN-Zero Module

We introduce a Self-Conditioned AdaLN-Zero (SCAdaLN-Zero) module to adapt ERP backbone features using the guidance feature produced by GAG. As shown in Fig. 4, unlike DiT-style AdaLN-Zero that uses external predefined 1D conditions such as timesteps or labels, our module uses an internal geometry-derived 2D signal from GAG. This guidance feature is used to predict the scale, shift, and gate parameters for the attention and MLP paths, rather than being directly fused into the backbone features.

This design is motivated by the need to adapt a perspective-pretrained backbone to panoramic inputs without disrupting its learned representation. The backbone already contains useful geometric priors from perspective image pretraining, while ERP inputs introduce projection distortion and distribution shift. A direct way to use cross-projection information is residual fusion or cross-attention. However, the ablation results in Table 3 show that such direct integration can degrade performance, suggesting that stable adaptation depends on how the cross-projection guidance is injected. Therefore, SCAdaLN-Zero uses the GAG feature only to generate normalization parameters, avoiding direct overwrite of the backbone feature content. Our design follows two key insights. First, scale and shift modulation provides a more controlled way to adapt pretrained features than explicit feature fusion, because it reweights normalized features instead of replacing their content. Second, the conditioning signal is generated inside the network from GAG, rather than from external prompts, labels, or timesteps. We therefore call the module self-conditioned AdaLN-Zero. With this design, the backbone can use geometry-aligned cross-projection cues to handle panoramic distortion while preserving the structure learned during pretraining.

**Modulation Parameter Generation.** Using the geometry-aware guidance signal $\mathbf{F}_{\mathrm{GAG}} \in \mathbb{R}^{1\times C\times H\times W}$ from the GAG module, we generate modulation

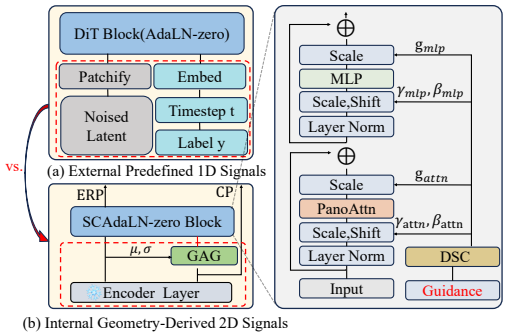


**Fig. 4:** The top-left panel illustrates the standard DiT block. The bottom-left panel shows our SCAdaLN-Zero extension embedded in the frozen backbone. The right side details the flow: it utilizes an internal geometry-derived signal from GAG as a self-conditioning signal to modulate backbone features.

parameters via a lightweight network comprising a SiLU activation followed by a depthwise separable convolution [9] (DSC):

$$\mathbf{P} = \mathrm{DSC}(\,\mathrm{SiLU}(\mathbf{F}_{\mathrm{GAG}})\,). \tag{7}$$

This architecture efficiently produces a parameter tensor $\mathbf{P} \in \mathbb{R}^{1\times 6C\times H\times W}$, which is subsequently split along the channel dimension into six independent parameter groups: $\boldsymbol{\beta}_{\mathrm{attn}}$, $\boldsymbol{\gamma}_{\mathrm{attn}}$, $\mathbf{g}_{\mathrm{attn}}$, $\boldsymbol{\beta}_{\mathrm{mlp}}$, $\boldsymbol{\gamma}_{\mathrm{mlp}}$, and $\mathbf{g}_{\mathrm{mlp}}$, where each group corresponds to modulation parameters for different components of the Transformer block.

These parameters facilitate fine-grained, implicit modulation across two critical paths in the Transformer block via a modulation function defined as:

$$\mathrm{modulate}(\mathbf{F}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{F} \odot (1 + \boldsymbol{\gamma}) + \boldsymbol{\beta}. \tag{8}$$

where $\mathbf{F}$ denotes the normalized input features, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ represent the adaptive scaling and shifting parameters, and $\odot$ indicates element-wise multiplication.

In the self-attention path, the input features are first normalized using a LayerNorm layer configured without its standard learnable affine parameters. This design ensures that the subsequent adaptive scaling and shifting serve as the sole affine transformation after normalization. The normalized features are then transformed by the modulate function (Eq. 8) using parameters $\boldsymbol{\beta}_{\mathrm{attn}}$ and $\boldsymbol{\gamma}_{\mathrm{attn}}$, processed by a panoramic self-attention [23] (PanoAttn) layer, and finally modulated by the gating parameter $\mathbf{g}_{\mathrm{attn}}$ to control the strength of the residual connection. In the modulated feed-forward network path, after applying LayerNorm to intermediate features and modulation via Eq. 8 with parameters $\boldsymbol{\beta}_{\mathrm{mlp}}$ and $\boldsymbol{\gamma}_{\mathrm{mlp}}$, we process them through an MLP layer, and finally gate them using $\mathbf{g}_{\mathrm{mlp}}$ to regulate the residual connection.

**Zero-Initialization Strategy.** Following DiT, we zero-initialize the final convolution in the modulation network, so the module initially degenerates to a standard Transformer block, ensuring stable training.

SCAdaLN-Zero reformulates panoramic adaptation as task-aligned parameter-space modulation rather than generic fine-tuning. By injecting geometry-aware cross-projection signals into normalization parameters, it enables distortion-aware refinement while preserving pretrained perspective priors.

### 3.3 Loss Function

To mitigate distortion-induced imbalance in ERP supervision, we introduce a cubemap-domain consistency constraint, termed the E2C Consistency Loss (ECCLoss). This design is motivated by the observation that spherical projection causes polar regions to occupy disproportionate pixel coverage in ERP, while informative geometric structures are often concentrated in equatorial areas. Such mismatch between pixel distribution and information density may bias supervision toward distortion-heavy regions and weaken depth learning in geometrically informative areas. We therefore apply depth constraints in the CP domain, where each face follows standard perspective geometry and naturally separates polar and equatorial regions. This formulation reduces the impact of spherical distortion present in ERP supervision and provides a perspective-consistent domain for learning depth structures in a more geometrically balanced manner.

The core formulation transforms both the predicted and ground truth depth maps from ERP to CP, where $\mathbf{D}^x$ and $\mathbf{D}$ denote the predicted and ground truth depth maps in cubemap format $\mathbb{R}^{6\times 1\times H_c\times W_c}$, respectively. The loss function employs a Scale-Shift Invariant Mean Absolute Error [12] (SSI-MAE) formulation:

$$\mathcal{L}_{\text{ECC}} = \frac{1}{B}\sum_{i=1}^{B} \left\|\psi(\mathbf{D}_i) - \psi(\mathbf{D}_i^x)\right\|_1 . \tag{9}$$

Here, $B$ denotes the number of depth maps used for loss computation, typically set to $B = 6$ corresponding to the six faces of the cubemap. The function $\psi(\cdot)$ represents the scale-shift normalization function:

$$\psi(\mathbf{D}) = \frac{\mathbf{D} - \mu(\mathbf{D})}{\sigma(\mathbf{D}) + \epsilon}. \tag{10}$$

The moments $\mu(\mathbf{D})$ and $\sigma(\mathbf{D})$ are computed over the full depth map by averaging over $(H, W)$ (not per channel):

$$\mu(\mathbf{D}) = \frac{1}{HW}\sum_{h,w} D_{h,w}, \quad \sigma(\mathbf{D}) = \sqrt{\frac{1}{HW}\sum_{h,w}\big(D_{h,w} - \mu(\mathbf{D})\big)^2 + \epsilon}. \tag{11}$$

with $\epsilon$ for numerical stability.

The geometric rationale of this approach stems from three advantages of CP: First, it naturally separates polar and equatorial regions through six planes, minimizing spherical distortion. Second, each face preserves perspective geometry, enabling learning of authentic 3D structure. Finally, this format reduces erroneous depth correlations between polar and equatorial regions common in ERP.

In addition to ECCLoss, our overall supervision incorporates two widely-used depth estimation losses: SILog loss $L_{\text{SILog}}$ [12] and gradient-domain loss $L_{\text{Grad}}$ [37]. The final supervised loss function is defined as:

$$\begin{aligned} L_S(\mathbf{D}^x, \mathbf{D}) = L_{\text{SILog}}(\mathbf{D}^x, \mathbf{D}) + L_{\text{Grad}}(\mathbf{D}^x, \mathbf{D}) \\ + \lambda_E \cdot L_{\text{ECC}}(\mathbf{D}^x, \mathbf{D}). \end{aligned} \tag{12}$$

where $\mathbf{D}^x$ and $\mathbf{D}$ denote the predicted and ground truth depth maps, with $\lambda_E$ balancing the ECCLoss contribution.

## 4 Experiment

### 4.1 Implementation Details

**Datasets.** We employ two real-world indoor panoramic datasets, Matterport3D [7] and Stanford2D3D [3], for evaluation under two settings: for in-domain evaluation, models are trained directly on these two datasets; for zero-shot evaluation, models are trained exclusively on synthetic data from Structured3D [42] and Deep360 [18]. Following PanDA's evaluation protocol, we train at $504 \times 1008$ and upsample all predictions to $512 \times 1024$ for evaluation.

**Implementation Details.** All experiments are performed on eight NVIDIA RTX A4000 GPUs using the PyTorch framework. We employ the Adam optimizer [17] with an initial learning rate of $5 \times 10^{-5}$, while keeping other hyperparameters at their default values. During training, we use a batch size of 1 and adopt several data augmentation strategies, including left-right flipping, panoramic horizontal rotation, and luminance adjustment, following established practices in [14, 23]. Our backbone network is the frozen Depth Anything Model v2 [36] (DAM v2), initialized with the same pretrained weights as PanDA [6]. To balance efficiency and computational cost, we integrate our proposed modules into the odd-numbered layers, with layer ablations provided in the supplementary material.

### 4.2 Comparison Analysis

**Quantitative Results.** As shown in Table 1, our method achieves SOTA performance on both Matterport3D [7] and Stanford2D3D [3] datasets. Notably, PanDA-L [6] first undergoes large-scale semi-supervised multi-dataset pretraining on four panoramic datasets (Structured3D [42], Deep360 [18], ZInD [10], and 360+x [8]), totaling about 120K samples, and is then fine-tuned on in-domain data. In contrast, our method surpasses it using only in-domain training data (1K for Stanford2D3D and 8K for Matterport3D), highlighting data efficiency. In a fair comparison with the non-pretrained PanDA-L, our advantages become more pronounced: improving Abs Rel by 12.3% and RMSE by 17.3% on Matterport3D, while achieving 34.2% and 22.3% improvements respectively on Stanford2D3D. The zero-shot generalization results in Table 2 further validate our advantages under the same synthetic-data training regime (Structured3D and Deep360 only), with improvements of 42.3% in Abs Rel and 14.0% in RMSE on Stanford2D3D.

**Table 1:** Quantitative comparison with SOTA methods on Matterport3D and Stanford2D3D datasets. The best results are in **bold**, and the second best are underlined. The $\Delta$ row indicates the relative improvement (in percentage) of our method over the previous best method (PanDA-L) under fair comparison settings. Note that the training protocols differ: PanDA-L* was first pre-trained using semi-supervised learning on 120K panoramic images and then fine-tuned on in-domain data, whereas our method and all other baselines were trained directly on in-domain data only.

| Dataset | Method | Abs Rel ↓ | Sq Rel ↓ | RMSE ↓ | $\delta_1$ ↑ | $\delta_2$ ↑ | $\delta_3$ ↑ |
|---|---|---|---|---|---|---|---|
| Matterport3D [7] | BiFuse [25] | 0.2048 | - | 0.6259 | 84.52 | 93.19 | 96.32 |
| | UniFuse [14] | 0.1063 | - | 0.4941 | 88.97 | 96.23 | 98.31 |
| | BiFuse++ [26] | - | - | 0.5190 | 87.90 | 95.17 | 97.72 |
| | PanoFormer [23] | - | - | 0.3635 | 91.84 | 98.04 | 99.16 |
| | HRDFuse [1] | 0.0967 | 0.0936 | 0.4433 | 91.62 | 96.69 | 98.44 |
| | Elite360D [2] | 0.1115 | 0.0914 | 0.4875 | 88.15 | 96.46 | 98.74 |
| | Depth Anywhere [32] | - | 0.0850 | - | 91.70 | 97.60 | 99.10 |
| | SGFormer [41] | 0.1039 | 0.0865 | 0.4790 | 89.46 | 96.42 | 98.59 |
| | PanDA-L* [6] | 0.0717 | - | 0.3305 | 95.09 | 98.94 | 99.65 |
| | PanDA-L | 0.0788 | 0.0572 | 0.3827 | 93.73 | 97.90 | 99.44 |
| | RePer-360 (Ours) | **0.0691** | **0.0388** | **0.3164** | **95.67** | **99.07** | **99.69** |
| | $\Delta$ | 12.3% | 32.2% | 17.3% | 2.07% | 1.19% | 0.25% |
| Stanford2D3D [3] | BiFuse [25] | 0.1209 | - | 0.4142 | 86.60 | 95.80 | 98.60 |
| | UniFuse [14] | 0.1114 | - | 0.3691 | 87.11 | 96.64 | 98.82 |
| | BiFuse++ [26] | - | 0.3720 | - | 87.83 | 96.49 | 98.84 |
| | PanoFormer [23] | - | 0.3083 | - | 93.94 | 98.38 | 99.41 |
| | HRDFuse [1] | 0.0935 | 0.0508 | 0.3106 | 91.40 | 97.98 | 99.27 |
| | Elite360D [2] | 0.1182 | 0.0728 | 0.3756 | 88.72 | 96.84 | 98.92 |
| | Depth Anywhere [32] | 0.1180 | - | 0.3510 | 91.00 | 97.10 | 98.70 |
| | SGFormer [41] | 0.1040 | 0.0581 | 0.3406 | 89.98 | 96.93 | 99.08 |
| | PanDA-L* [6] | 0.0609 | - | 0.2540 | 96.82 | **99.05** | **99.52** |
| | PanDA-L | 0.0881 | 0.0550 | 0.3185 | 93.42 | 98.26 | 99.33 |
| | RePer-360 (Ours) | **0.0580** | **0.0382** | **0.2474** | **97.37** | 98.89 | 99.35 |
| | $\Delta$ | 34.2% | 30.5% | 22.3% | 4.22% | 0.64% | 0.02% |

These results demonstrate cross-domain generalization and data efficiency. See the supplementary material for complexity analysis and comparison with ST$^2$360D.

**Qualitative Results.** As shown in Fig. 5, our method consistently outperforms PanDA-L [6]. PanDA-L tends to misinterpret wall textures as depth variations and loses structural details in complex regions, whereas our approach preserves scene geometry and fine structures more faithfully, especially under severe panoramic distortions. These visual improvements align with our superior RMSE results, demonstrating effective distortion handling despite using substantially less training data than PanDA-L. Fig. 6 further shows zero-shot results on SUN360. Trained only on synthetic panoramas (Structured3D and Deep360), our model achieves stable geometric reconstruction under challenging outdoor illumination, while PanDA-L is more sensitive to lighting changes and occasionally confuses them with depth variations. These results highlight the strong cross-domain generalization of our method with limited panoramic data. Additional results on self-collected real-world panoramas are provided in the supplementary material.

**Table 2:** Comparison of zero-shot depth estimation performance. The best results are in **bold**.

| Method | Backbone | Matterport3D [7] | | Stanford2D3D [3] | |
|---|---|---|---|---|---|
| | | Abs Rel ↓ | RMSE ↓ | Abs Rel ↓ | RMSE ↓ |
| Marigold [16] | SD 2.0 [22] | - | 0.5745 | - | 0.5069 |
| DAM v1 [36] | ViT-L | - | 1.1431 | - | 0.7597 |
| DAM v2 [36] | ViT-L | - | 0.5522 | - | 0.4884 |
| PanDA-L [6] | ViT-L | 0.1036 | 0.4539 | 0.1092 | 0.3314 |
| RePer-360 (Ours) | ViT-L | **0.1033** | **0.4534** | **0.0630** | **0.2849** |

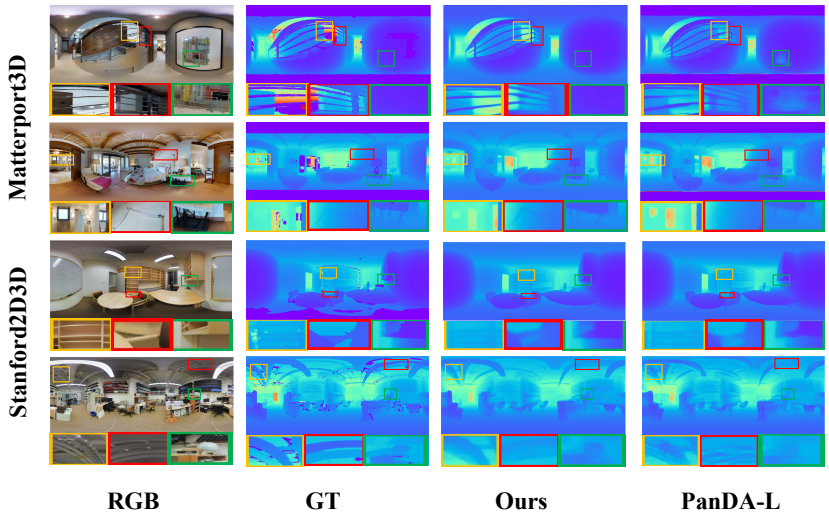


**Fig. 5:** Qualitative comparison between the current SOTA method PanDA-L [6] and ours. The results of PanDA-L are obtained using their released weights. The top sample is from Matterport3D, while the bottom one is from Stanford2D3D.

### 4.3 Architectural Comparison and Ablation

We conduct comprehensive architectural analyses on Matterport3D [7] under a unified training configuration. All variants share the same frozen backbone, training data, and optimization strategy to ensure fairness. In addition to internal structural ablations, we further compare with representative traditional panoramic fusion designs to evaluate the effectiveness of geometry-aware modulation. Results are summarized in Table 3.

**Effectiveness of SCAdaLN-Zero.** As shown in Table 3 (Index a–d), replacing explicit interaction (Index a/b) with SCAdaLN-Zero (Index c/d) yields clear improvements. Even single-branch conditioning (ERP or CP) outperforms the degenerated baseline, while the full geometry-aligned configuration (Index d)

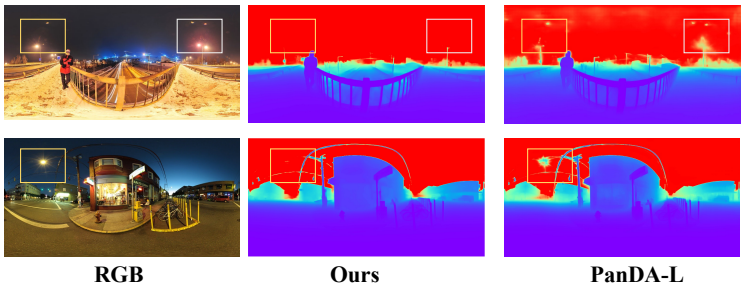


**Fig. 6:** Zero-shot qualitative comparison on the unlabeled SUN360 [34] dataset. Our model is trained only on Structured3D [42] and Deep360 [18], while PanDA-L [6] uses substantially more panoramic supervision. Despite the limited training data, our method better preserves structural geometry under complex outdoor illumination.

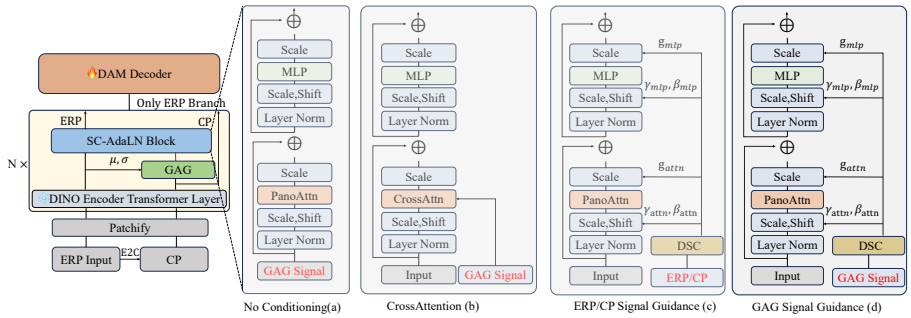


**Fig. 7: Architectural Ablation Study.** We compare different structural formulations to validate our design: (a) a degenerated version without conditioning; (b) replacing normalization-based modulation with explicit feature mixing via cross-attention; (c) conditioning on a single-branch signal (ERP or CP); and (d) the complete geometry-aware conditioning design (ours), which achieves the best performance.

performs best. These results show that normalization-based modulation is more effective than explicit value-level feature mixing for cross-domain adaptation.

**Effect of Geometry-Aligned Guidance.** Comparing Index c (ERP/CP single-branch variants) with Index d, incorporating GAG conditioning further improves performance over single-branch signals. This indicates that geometry-aligned cross-projection features provide more informative and consistent modulation cues than using isolated ERP or CP features alone.

**Fusion Architectural Variants.** We evaluate two variants using the CEE module from UniFuse: replacing GAG with CEE fusion while retaining SCAdaLN, and adopting a full fusion design without modulation. The former underperforms geometry-aligned conditioning, while the latter incurs a substantial drop, indicating that multi-branch fusion is insufficient for projection-gap adaptation.

**Table 3: Architectural Ablation Results.** Quantitative comparison of different architectural variants of the proposed framework, including alternative formulations of the SCAdaLN-based conditioning mechanism, attention types, and the effect of ECCLoss. Configurations (a–d) correspond to Fig. 7. Best results are shown in **bold**.

| Index | SCAdaLN | Guidance Source | Attention | ECCLoss | RMSE↓ | Abs Rel↓ | Sq Rel↓ | $\delta_1$(%) ↑ | $\delta_2$(%) ↑ |
|---|---|---|---|---|---|---|---|---|---|
| - | ✗ | None | None | ✗ | 0.3846 | 0.0915 | 0.0525 | 92.25 | 98.53 |
| a | ✗ | GAG | PanoAttn | ✓ | 0.4638 | 0.1158 | 0.0790 | 86.80 | 96.80 |
| b | ✗ | GAG | CrossAttn | ✓ | 0.4549 | 0.1168 | 0.0763 | 87.15 | 97.17 |
| c | ✓ | ERP | PanoAttn | ✓ | 0.3206 | 0.0698 | 0.0390 | 95.66 | 99.03 |
| c | ✓ | CP | PanoAttn | ✓ | 0.3227 | 0.0692 | **0.0388** | 95.58 | 99.03 |
| d | ✓ | GAG | PanoAttn | ✓ | **0.3164** | **0.0691** | **0.0388** | **95.67** | **99.07** |
| - | ✓ | GAG | PanoAttn | ✗ | 0.3249 | 0.0716 | 0.0396 | 95.17 | 99.04 |
| - | ✓ | GAG | NormalAttn | ✓ | 0.3209 | 0.0725 | 0.0393 | 95.43 | 99.01 |
| - | ✓ | UniFuse | PanoAttn | ✓ | 0.3244 | 0.0731 | 0.0412 | 94.93 | 99.04 |
| - | ✗ | UniFuse | None | ✓ | 0.6032 | 0.1641 | 0.1426 | 78.30 | 93.73 |

**Table 4:** Impact of Pretrained Initialization on Matterport3D (ViT-L backbone).

| Method | AbsRel ↓ | RMSE ↓ | $\delta_1$(%) ↑ |
|---|---|---|---|
| PanDA (DAMv2) | 0.0788 | 0.3827 | 93.73 |
| Ours (DAMv2) | **0.0691** | **0.3164** | **95.67** |
| Ours (DINOv2) | 0.0808 | 0.3518 | 94.33 |

**Role of ECCLoss and Attention Design.** As shown in the lower rows of Table 3, removing ECCLoss leads to consistent performance degradation, highlighting its importance in enforcing cross-projection geometric consistency. Replacing panoramic attention with standard self-attention produces only minor performance differences, indicating that the proposed framework is not tightly coupled to a specific attention operator.

**Impact of Pretrained Initialization.** We evaluate the role of pretrained weights using the same ViT-L architecture (Table 4). With identical DAMv2 initialization, our method consistently outperforms the LoRA-based PanDA baseline. When initialized from task-agnostic DINOv2 weights, RePer-360 remains competitive, suggesting that the proposed adaptation is not solely dependent on depth-pretrained initialization.

### 4.4 Representation Drift and Feature Visualization

To better understand different adaptation mechanisms, we analyze representation drift and feature-level responses.

**Representation Drift.** The left panel of Fig. 8 reports feature similarity to the frozen backbone on Matterport3D using cosine similarity and linear CKA. Cross-attention induces large representational shifts with unstable layer-wise trends. In contrast, PanDA stays close to the backbone, while RePer-360 preserves high similarity with smooth, consistent inter-layer evolution, indicating controlled, geometry-aware adaptation rather than layer-wise instability. Overall, effective

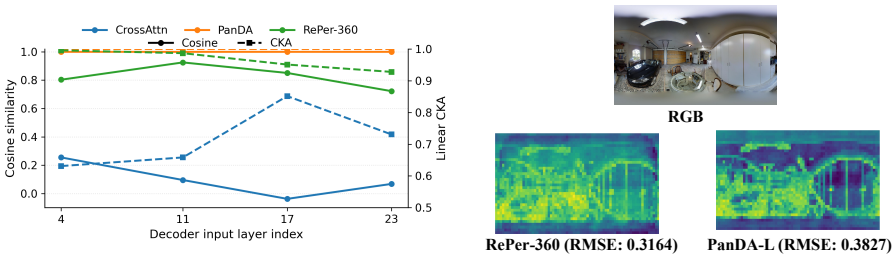


**Fig. 8: Left:** Feature drift relative to the frozen backbone on Matterport3D (the first 500 test images), measured using cosine similarity and linear CKA across transformer layers. **Right:** Visualization of layer-4 (decoder-input) features for an example image, comparing RePer-360 and the in-domain PanDA baseline. Both methods capture a similar near-to-far transition, while PanDA shows over-activation on ceiling/lighting regions and an exaggerated ceiling–wall contrast.

panoramic adaptation benefits from balanced representation reconfiguration instead of excessive perturbation or near-identity preservation.

**Feature Map Visualization.** We visualize decoder-input features at layer 4 to compare PanDA and RePer-360. As shown in Fig. 8, both methods exhibit a consistent near-to-far intensity transition, indicating that the depth/layout trend is captured. However, PanDA shows disproportionately strong responses on ceiling regions and lighting fixtures, resulting in an exaggerated ceiling–wall contrast. In contrast, RePer-360 maintains a more coherent response on these planar regions, consistent with the controlled drift trends in the left panel.

## 5 Conclusion

We introduce RePer-360 to bridge the projection gap between perspective and panoramic imagery through geometry-aware, self-conditioned normalization-based modulation. By preserving and reorganizing pretrained perspective priors rather than overwriting them via feature fusion, our framework achieves structured cross-domain adaptation with strong data efficiency. These findings suggest that balanced prior modulation offers a principled strategy for adapting pretrained visual backbones to geometrically mismatched domains.

***Acknowledgement.*** *This work was supported by the National Natural Science Foundation of China (Nos. 62573039, U2441242).*

## References

1. Ai, H., Cao, Z., Cao, Y.P., Shan, Y., Wang, L.: Hrdfuse: Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 13273–13282 (2023), `https://api.semanticscholar.org/CorpusID:257636739`

2. Ai, H., Wang, L.: 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 (CVPR). pp. 9926–9935 (June 2024)
3. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. CoRR **abs/1702.01105** (2017)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
5. Cao, Z., Zhu, J., Ai, H., Jiang, L., Lyu, Y., Xiong, H.: St$^2$360d: Spatial-to-temporal consistency for training-free 360 monocular depth estimation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
6. Cao, Z., Zhu, J., Zhang, W., Ai, H., Bai, H., Zhao, H., Wang, L.: Panda: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 982–992 (2025)
7. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: 3DV (2017)
8. Chen, H., Hou, Y., Qu, C., Testini, I., Hong, X., Jiao, J.: 360+ x: A panoptic multi-modal scene understanding dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19373–19382 (2024)
9. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
10. Cruz, S., Hutchcroft, W., Li, Y., Khosravan, N., Boyadzhiev, I., Kang, S.B.: Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2133–2143 (2021)
11. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
12. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems (2014)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
14. Jiang, H., Sheng, Z., Zhu, S., Dong, Z., Huang, R.: Unifuse: Unidirectional fusion for 360° panorama depth estimation. IEEE Robotics and Automation Letters **6**, 1519–1526 (2021), `https://api.semanticscholar.org/CorpusID:231846497`
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
16. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9492–9502 (2024)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Li, M., Jin, X., Hu, X., Dai, J., Du, S., Li, Y.: Mode: Multi-view omnidirectional depth estimation with 360° cameras. In: European Conference on Computer Vision. pp. 197–213. Springer (2022)

19. Li, Y.y., Guo, Y., Yan, Z., Huang, X., Duan, Y., Ren, L.: Omnifusion: 360 monocular depth estimation via geometry-aware fusion. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2791–2800 (2022), https://api.semanticscholar.org/CorpusID:247218344
20. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
21. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
23. Shen, Z., Lin, C., Liao, K., Nie, L., Zheng, Z., Zhao, Y.: Panoformer: Panorama transformer for indoor 360° depth estimation. In: European Conference on Computer Vision (2022), https://api.semanticscholar.org/CorpusID:247519181
24. Shen, Z., Lin, C., Nie, L., Liao, K., Lin, W., Zhao, Y.: Revisiting 360 depth estimation with panogabor: A new fusion perspective. arXiv preprint arXiv:2408.16227 (2024)
25. Wang, F.E., Yeh, Y.H., Sun, M., Chiu, W.C., Tsai, Y.H.: Bifuse: Monocular 360 depth estimation via bi-projection fusion. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 459–468 (2020), https://api.semanticscholar.org/CorpusID:219963198
26. Wang, F.E., Yeh, Y.H., Tsai, Y.H., Chiu, W.C., Sun, M.: Bifuse++: Self-supervised and efficient bi-projection fusion for 360° depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**, 5448–5460 (2022), https://api.semanticscholar.org/CorpusID:252010287
27. Wang, J., Lin, C., Guan, C., Nie, L., He, J., Li, H., Liao, K., Zhao, Y.: Jasmine: Harnessing diffusion prior for self-supervised depth estimation. arXiv preprint arXiv:2503.15905 (2025)
28. Wang, J., Lin, C., Nie, L., Liao, K., Shao, S., Zhao, Y.: Digging into contrastive learning for robust depth estimation with diffusion models. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 4129–4137. MM '24, ACM (Oct 2024). https://doi.org/10.1145/3664647.3681168, http://dx.doi.org/10.1145/3664647.3681168
29. Wang, J., Lin, C., Sun, L., Cao, Z., Yin, Y., Nie, L., Yuan, Z., Chu, X., Wei, Y., Liao, K., et al.: Geometry-guided reinforcement learning for multi-view consistent 3d scene editing. arXiv preprint arXiv:2603.03143 (2026)
30. Wang, J., Lin, C., Sun, L., Liu, R., Nie, L., Li, M., Liao, K., Chu, X., Zhao, Y.: From editor to dense geometry estimator. arXiv preprint arXiv:2509.04338 (2025)
31. Wang, J., Ouyang, H., Lin, J., Lin, C., Fan, D., Zhang, B., Fan, H., Zuo, F., Sun, J., Wang, H., et al.: Cac: Advancing video reward models via hierarchical spatiotemporal concentrating. arXiv preprint arXiv:2605.11723 (2026)
32. Wang, N.H.A., Liu, Y.L.: Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. Advances in Neural Information Processing Systems **37**, 127739–127764 (2024)
33. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. arXiv preprint arXiv:2507.02546 (2025)
34. Xiao, J., Ehinger, K.A., Oliva, A., Torralba, A.: Recognizing scene viewpoint using panoramic place representation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2695–2702. IEEE (2012)

35. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
36. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
37. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9043–9053 (2023)
38. Yoon, Y., Chung, I., Wang, L., Yoon, K.J.: Spheresr: 360° image super-resolution with arbitrary projection via continuous spherical image representation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5667–5676 (2021)
39. Yun, I., Shin, C.Y., Lee, H., Lee, H.J., Rhee, C.E.: Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. ArXiv **abs/2304.07803** (2023)
40. Zhang, J., Yang, K., Ma, C., Reiß, S., Peng, K., Stiefelhagen, R.: Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16917–16927 (2022)
41. Zhang, J., Chen, Z., Lin, C., Shen, Z., Nie, L., Liao, K., Zhao, Y.: Sgformer: Spherical geometry transformer for 360 depth estimation. IEEE Transactions on Circuits and Systems for Video Technology (2025)
42. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: European Conference on Computer Vision (2019), `https://api.semanticscholar.org/CorpusID:199064623`
43. Zheng, X., Pan, T., Luo, Y., Wang, L.: Look at the neighbor: Distortion-aware unsupervised domain adaptation for panoramic semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18687–18698 (2023)
44. Zheng, X., Zhu, J., Liu, Y., Cao, Z., Fu, C., Wang, L.: Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1285–1295 (2023)