SO(3)-Equivariant ViT-Adapter for Data-Efficient Zero-Shot Sim-to-Real Indoor Panoramic Depth Estimation Supplemental Material

A. Theoretical Details of the Equivariant Spherical Prior Module

To understand how the proposed Equivariant Spherical Prior Module (ESPM) preserves rotation-equivariance, we begin by revisiting the concept of group equivariance and its connection to spherical convolution.

Group Equivariance. Let $\phi : V_{in} \to V_{out}$ denote a mapping between input and output feature spaces. The map ϕ is defined to be *equivariant* with respect to an abstract group G, if for all $g \in G$ there exist transformations $T_g : V_{in} \to V_{in}$ and $S_g : V_{out} \to V_{out}$ such that:

$$S_g[\phi(v)] = \phi(T_g[v]), \tag{1}$$

In other words, transforming the input and then applying ϕ yields the same result as applying ϕ and then transforming the output.

Equivariance in Planar Convolution. For perspective images defined on the Euclidean plane, consider a conventional 2D convolution layer $\phi(f) = \kappa * f$ with kernel κ. The translation operator L_t shifts the input feature map f by $t = (t_x, t_y)$, such that:

$$(L_t f)(x, y) = f(x - t_x,\, y - t_y), \tag{2}$$

or in group action notation, $L_t f(p) = f(t^{-1}p)$, where p denotes a point on $\mathbb{R}^2$. From the definition of convolution, it follows that:

$$(\kappa * f)(p) = \int_{\mathbb{R}^2} \kappa(q)\, f(p - q)\, dq, \tag{3}$$

Applying the translation operator $L_t f(p) = f(p - t)$ gives:

$$\begin{aligned}(\kappa * (L_t f))(p) &= \int_{\mathbb{R}^2} \kappa(q)\, f((p - q) - t)\, dq \\ &= (L_t(\kappa * f))(p),\end{aligned} \tag{4}$$

Therefore, $L_t(\kappa * f) = \kappa * (L_t f)$, demonstrating that standard convolution layers are inherently translation-equivariant. Consequently, CNNs operating on planar images naturally encode local spatial structures and translation-consistent priors, which can be leveraged as inductive biases in vision transformers [6, 8].

From Planar to Spherical Representation. Unlike perspective images, panoramic images are parameterized by spherical coordinates (α, β), where $\alpha \in [0, 2\pi)$ and $\beta \in [0, \pi]$. A "translation" on this spherical grid no longer corresponds to a planar translation, but instead to a rotation in 3D space. Each pixel movement on the panorama can be interpreted as rotating its associated unit vector on the sphere. Hence, for panoramic images, it is necessary to introduce a rotation operator L_R to replace L_t on perspective images:

$$(L_R f)(x) = f(R^{-1}x), \quad R \in \text{SO(3)}. \tag{5}$$

Rotation Group and Homogeneous Space. All possible 3D rotations form the special orthogonal group SO(3), where each rotation element R can be represented as a 3×3 unit orthogonal matrix. To describe rotations more intuitively, a compact parameterization can be obtained using ZYZ–Euler angles:

$$R(\alpha, \beta, \gamma) = R_z(\alpha) R_y(\beta) R_z(\gamma), \tag{6}$$

where $\alpha, \gamma \in [0, 2\pi]$ and $\beta \in [0, \pi]$, with R_z and R_y denoting rotations about the Z and Y axes, respectively. Similarly, a point $x \in S^2$ can be expressed as:

$$x(\alpha, \beta) = R_z(\alpha) R_y(\beta)\mathbf{n}, \tag{7}$$

where $\mathbf{n}$ denotes the north pole. This formulation reveals that any point on the sphere can be generated by applying a rotation from SO(3) to the north pole. Consequently, the sphere S^2 is isomorphic to the quotient group SO(3)/SO(2), meaning it can be regarded as a homogeneous space of SO(3). The corresponding invariant subgroup of SO(3) is SO(2), and its elements can be parameterized as $R_z(\gamma)$.

Spherical Convolution. To construct rotation-equivariant representations on the sphere, we need to define convolution operations that are consistent with the topology of S^2. However, unlike S^2, when the Euclidean plane $\mathbb{R}^2$ is considered

as a homogeneous space of the 2D translation group T(2), the corresponding invariant subgroup is trivial. This distinction implies that when performing rotation-equivariant convolutions on spherical signals $f : \mathrm{S}^2 \to \mathbb{R}^c$, the resulting features are lifted to SO(3), and subsequent convolutions will continue to operate on SO(3). Specifically, taking the spherical function $f : \mathrm{S}^2 \to \mathbb{R}^{c_{in}}$ as input, the first layer S2Conv with kernel $\kappa : \mathrm{S}^2 \to \mathbb{R}^{c_{out} \times c_{in}}$ is defined by the inner product, which lifts f to an SO(3) function $h : \mathrm{SO}(3) \to \mathbb{R}^{c_{out}}$:

$$\begin{aligned} h(R) = (\kappa * f)(R) &=< L_R\kappa, f > \\ &= \int_{\mathrm{S}^2} \kappa(R^{-1}x) f(x) dx, \end{aligned} \tag{8}$$

where $dx = d\alpha d\beta sin(\beta)$ is the Haar measure on the sphere. Subsequently, SO3Conv with a kernel $\kappa : \mathrm{SO}(3) \to \mathbb{R}^{c_{out} \times c_{in}}$ is defined in a similar way, taking the SO(3) function $h : \mathrm{SO}(3) \to \mathbb{R}^{c_{in}}$ as input:

$$\begin{aligned} h'(R) = (\kappa * h)(R) &=< L_R\kappa, h > \\ &= \int_{\mathrm{SO(3)}} \kappa(R^{-1}Q) h(Q) dQ, \end{aligned} \tag{9}$$

where $dQ = d\alpha d\beta sin(\beta) d\gamma$ is the Haar measure on SO(3). We employ S2Conv and SO3Conv introduced in [7], which together form the computational backbone of our Equivariant Spherical Prior Module (ESPM).

Direct computation of integrals on S^2 and SO(3) is computationally expensive. Non-uniform angular sampling can also cause misalignment between the kernels and functions after rotation. To address this problem, a common strategy is to compute convolutions in the spectral domain using the generalized Fourier transform. This approach is based on the Fourier convolution theorem, which states that a convolution in the spatial domain can be computed by pointwise multiplication of the corresponding coefficients in the Fourier domain:

$$\begin{aligned} \mathcal{F}(f * \kappa)(m) &= (\mathcal{F}f)(m)(\mathcal{F}\kappa)(m) \\ &= \hat{f}(m)\hat{\kappa}(m), \end{aligned} \tag{10}$$

where $\mathcal{F}(\cdot)$ denotes the generalized Fourier transform, and $\hat{f}(m) = \langle f, \psi_m \rangle$, $\hat{\kappa}(m) = \langle \kappa, \psi_m \rangle$ are the Fourier coefficients with respect to the orthonormal basis $\{\psi_m\}$ indexed by m.

Another notable advantage of this spectral formulation is that it allows exact computation under certain conditions. For compact homogeneous manifolds such as S^2 and SO(3), the Fourier domain is discrete, and band-limited functions can be represented by a finite number of coefficients. The sampling theorem ensures that when a signal is band-limited to a given bandwidth B, its harmonic coefficients can be computed exactly using quadrature weights. Therefore, the convolution obtained by multiplying the spectral coefficients and applying the inverse transform provides an exact computation of the continuous convolution within the chosen bandwidth. Specifically, for a band-limited spherical signal $f(\theta, \phi)$, the generalized Fourier transform expands f onto spherical harmonics Y_m^l:

$$\begin{aligned} f(\theta,\phi) &= \sum_{l=0}^{B-1} \sum_{m=-l}^{l} \hat{f}_m^l Y_m^l(\theta,\phi), \\ (\mathcal{F}f)_m^l &= \hat{f}_m^l = \sum_{i=0}^{2B-1} \sum_{j=0}^{2B-1} s(\theta,\phi), \\ s(\theta,\phi) &= f(\theta_i,\phi_j)\overline{Y_m^l(\theta_i,\phi_j)} q(\theta,\phi), \end{aligned} \tag{11}$$

where $q(\theta, \phi)$ are the quadrature weights. For a function $f(\alpha, \beta, \gamma)$ on SO(3), we use the Wigner–D basis D_l^{mn}:

$$\begin{aligned} f(\alpha,\beta,\gamma) &= \sum_{l=0}^{B-1} \frac{2l+1}{8\pi^2} \sum_{m=-l}^{l} \sum_{n=-l}^{l} \hat{f}_{mn}^l \overline{D_{mn}^l(\alpha,\beta,\gamma)}, \\ (\mathcal{F}f)_{mn}^l &= \hat{f}_{mn}^l = \sum_{i=0}^{2B-1} \sum_{j=0}^{2B-1} \sum_{k=0}^{2B-1} s(i,j,k) \\ s(i,j,k) &= f(\alpha_i,\beta_j,\gamma_k) D_{mn}^l(\alpha_i,\beta_j,\gamma_k) q(\alpha_i,\beta_j,\gamma_k) \end{aligned}$$

where $q(\alpha_i, \beta_j, \gamma_k)$ are the quadrature weights corresponding to the SO(3) measure.

B. Additional Experiments under Resource-Rich Training Regimes

While the main paper focuses on the challenging zero-shot sim-to-real setting, we additionally report experimental results under more conventional resource-rich training configurations commonly adopted in panoramic depth estimation. All results are evaluated with median alignment and standard pixel-wise metrics without spherical correction, following the standard protocol used in prior panoramic depth estimation works. Unless otherwise noted, numbers for competing methods are directly quoted from their original publications.

B.1. Comparison with SOTA Panoramic Methods

To enable a broader comparison with existing panoramic depth estimation systems, we provide in-domain results on the Matterport3D dataset. We train our SO(3)-Equivariant ViT-Adapter and DPT head on the Matterport3D train split and evaluate on its official test split. The compared methods include the standard in-domain panoramic models and the zero-shot panoramic models with in-domain fine-tuning.

Table 1 summarizes the results. Our SO(3)-Equivariant ViT-Adapter with a ViT-L backbone achieves the best performance across almost all metrics, surpassing both in-domain

Method	Train Data	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
BiFuse [14]	*MP3D*	/	/	0.626	0.845	0.932	0.963
UniFuse [9]	*MP3D*	0.106	/	0.494	0.890	0.962	0.983
HoHoNet [13]	*MP3D*	/	/	0.514	0.879	0.952	0.977
BiFuse++ [15]	*MP3D*	/	/	0.519	0.879	0.952	0.977
PanoFormer [12]	*MP3D*	/	/	0.364	0.918	0.980	0.992
OmniFusion [10]	*MP3D*	0.090	0.055	0.426	0.919	0.980	0.993
HRDFuse [1]	*MP3D*	0.097	0.094	0.443	0.916	0.967	0.984
SphereFusion [18]	*MP3D*	/	/	0.489	0.870	0.961	0.984
Elite360D [2]	*MP3D*	0.112	0.091	0.488	0.882	0.965	0.987
SGFormer [20]	*MP3D*	0.104	0.087	0.479	0.895	0.964	0.986
DepthAnywhere [16]	*MP3D, ST(p)*	0.085	/	/	0.917	0.976	0.991
PanDA-S [4]	*Mixed-datasets, MP3D*	0.092	/	0.395	0.923	0.983	0.995
PanDA-B [4]	*Mixed-datasets, MP3D*	0.079	/	0.348	0.946	0.988	0.996
PanDA-L [4]	*Mixed-datasets, MP3D*	0.072	/	0.331	0.951	0.989	**0.997**
Ours (ViT-S)	*Pers.(pt), MP3D*	0.074	0.048	0.363	0.944	0.987	0.995
Ours (ViT-B)	*Pers.(pt), MP3D*	0.070	0.043	0.340	0.952	0.989	0.996
Ours (ViT-L)	*Pers.(pt), MP3D*	**0.067**	**0.039**	**0.325**	**0.957**	**0.990**	0.996

Table 1. Quantitative comparison on the Matterport3D dataset. **Train Data** lists the datasets employed during training for the model used in inference: *MP3D* refers to the training split of the Matterport3D dataset; *Pers.(pt)* denotes the large-scale perspective datasets used during pre-training; *ST(p)* indicates the unlabeled Structured3D dataset where pseudo-labels are generated by a teacher model; and *Mixed-datasets* refers to the panoramic mixed datasets used by PanDA, consisting of labeled synthetic datasets and unlabeled real-world datasets. Best results are highlighted in **bold**.

Method	Backbone	Train Data	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
DepthAnywhere	UniFuse	*MP3D, ST(p)*	0.086	/	/	0.924	0.977	0.990
	BiFuse++	*MP3D, ST(p)*	**0.082**	/	/	0.931	0.979	0.991
	HoHoNet	*MP3D, ST(p)*	0.088	/	/	0.920	0.979	**0.992**
	EGFormer	*MP3D, ST(p)*	0.086	/	/	0.923	0.976	0.990
Ours	ViT-S	*Pers.(pt), MP3D*	0.088	0.081	0.373	0.943	**0.982**	**0.992**
	ViT-B	*Pers.(pt), MP3D*	0.083	0.076	**0.355**	0.950	**0.982**	0.990
	ViT-L	*Pers.(pt), MP3D*	**0.082**	**0.072**	0.357	**0.953**	**0.982**	0.991

Table 2. Zero-shot real-to-real performance on the Stanford2D3D dataset. All other settings and notations are identical to Table 1.

baselines and zero-shot models with in-domain fine-tuning. Even our ViT-S adapter consistently outperforms most competing methods. Its performance is slightly lower only than the larger PanDA models that use ViT-B or ViT-L backbones, whereas it exceeds the PanDA-S and other panoramic depth estimation baselines. For example, compared with the in-domain method OmniFusion, our ViT-S adapter improves δ_1 from 0.919 to 0.944. When compared with the fine-tuned foundation model PanDA-L, our ViT-L adapter reduces AbsRel from 0.072 to 0.067. These results demonstrate that our adapter is not only a data-efficient zero-shot solution, but also an effective mechanism for transferring perspective weights to the panoramic depth estimation task of the target scene with in-domain training data.

B.2. Zero-Shot Real-to-Real Evaluation

DepthAnywhere demonstrates strong performance in zero-shot real-to-real transfer. To compare under the same protocol, we train our adapter and DPT head on Matterport3D and directly evaluate on the Stanford2D3D test split without any fine-tuning. As shown in Table 2, our ViT-S adapter already surpasses DepthAnywhere across all accuracy metrics. Our ViT-L adapter further improves these results, achieving the lowest AbsRel and raising the δ_1 score to 0.953, a significant improvement over the 0.931 achieved by DepthAnywhere with BiFuse++.

However, comparing Table 2 with the results in the main paper, we observe that training on Matterport3D and testing on Stanford2D3D yields lower performance than training on PNVS and testing on Stanford2D3D. A plausible explanation is the significant difference in depth distribution between Matterport3D and Stanford2D3D, while our method

uses only a simple BerHu loss without any normalization or alignment during training.

B.3. Sim-to-Real Generalization in Zero-Shot Panoramic Transfer

Recent zero-shot panoramic depth estimation methods typically rely on transferring perspective models to the panoramic domain. In such approaches, most of the depth priors are derived from networks pre-trained on large-scale perspective images, and are applied to panoramic inputs through a transfer mechanism.

A comparison between the results reported in this section and the sim-to-real experiments in the main paper reveals an interesting pattern. Existing zero-shot approaches, such as PanDA [4] and DepthAnywhere [16], achieve strong performance when real training data is available. However, their performance drops noticeably when the transfer is learned only from synthetic panoramas. In contrast, our method maintains state-of-the-art performance on real benchmarks even when the transfer is conducted purely with synthetic data.

This behavior can be understood from the perspective of transfer mechanisms. ViTs pre-trained on large-scale perspective datasets inherently possess strong sim-to-real generalization ability. However, transferring these priors to panoramic depth estimation is challenging due to the fundamental geometric discrepancy between perspective and panoramic representations. To address this mismatch, previous works such as PanDA [4] and DepthAnywhere [16] adopt distillation-based frameworks that provide large adaptation flexibility. While effective for learning the new geometry, such designs may weaken the direct utilization of the original backbone representations.

In contrast, classical adapter-based transfer preserves the pre-trained weights and thus retains the rich priors learned in the source domain. However, the limited adaptation capacity of standard adapters makes it difficult to capture the geometric variations introduced by panoramic projections. The proposed SO(3)-Equivariant ViT-Adapter addresses this issue by explicitly injecting equivariant priors, enabling the adaptation space to better cover the geometric discrepancy between perspective and panoramic domains. This design allows the model to more effectively exploit the inherited priors of the pre-trained ViT and thereby achieve strong zero-shot sim-to-real transfer even when the transfer is learned purely from synthetic panoramas.

C. Experimental Setup

C.1. Datasets

Matterport3D Dataset [5]. Matterport3D is a real-world RGB-D dataset collected using a multi-camera rig that captures 18 RGB-D perspective images at each viewpoint, which are stitched into a 1024×2048 equirectangular panorama with an effective field-of-view of 3.75 sr. The dataset contains 10800 panoramic RGB-D views from 90 indoor environments. Following the official split, 7829 panoramas from 61 scenes are used for training, 957 panoramas from 11 scenes for validation, and 2014 panoramas from 18 scenes for testing. We set the maximum depth to 20m during training and 10m during evaluation.

Stanford2D3D Dataset [3]. Stanford2D3D is another real-world RGB-D dataset captured using the same hardware as Matterport3D. Six indoor areas are reconstructed in 3D, and panoramic RGB-D images of resolution 1024×2048 are rendered from virtual camera viewpoints. The dataset contains 1413 panoramas and provides three official cross-validation splits. Following common practice, we adopt the first split, using 373 panoramas from area 5 for testing. For evaluation, we mask out the top and bottom 15.625% of pixels and set the maximum depth to 10m.

Structured3D Dataset [21]. Structured3D is a large-scale synthetic dataset comprising photorealistic indoor panoramas with rich 3D structural annotations, including dense depth. Virtual cameras are placed in 21,835 rooms across 3,500 house models, and each viewpoint is rendered under three furniture layouts and three lighting conditions at a resolution of 512×1024. The dataset contains more than 196K panoramic images with corresponding dense depth maps. After filtering invalid samples, we select 18029 panoramas with full furniture and default lighting to provide additional training data for DepthAnywhere and PanDA. The maximum depth is set to 10m.

PNVS Dataset [17]. The PNVS dataset is constructed from Structured3D and provides synthetic multi-view panoramic RGB-D sequences. For each source view, three target views of resolution 512×1024 are rendered and low-quality samples are removed. According to inter-camera spacing, PNVS is split into PNVS-easy (0.2–0.3m) and PNVS-hard (1.0–2.0m). PNVS-hard contains 8758 source views and 19940 rendered target views. After filtering invalid samples, we use 6548 source views from PNVS-hard as training data. We do not use multi-view information; instead, we treat PNVS-hard purely as a limited-size labeled synthetic dataset to evaluate the data efficiency and zero-shot sim-to-real generalization of our approach. The maximum depth is set to 10m during training.

C.2. Evaluation Metrics

Panoramic depth estimation is commonly evaluated under the ERP representation using absolute relative error (Abs Rel), squared relative error (Sq Rel), root mean squared error (RMSE), logarithmic RMSE (RMSE(log_{10})), and accuracy thresholds (δ). Due to the latitude-dependent distortion inherent in ERP, following prior work [22], we apply a latitude-

aware weighting when computing error-based metrics. For accuracy, we uniformly sample points on the sphere using the S^2 generalized spiral scheme [11] with $N' = 0.25 \times H \times W$ samples. The weighted error metrics are defined as:

$$
\begin{aligned}
AbsRel &= \frac{1}{|N|}\sum_{i\in N}\frac{|d_i - d_i^*|}{d_i^*}\cdot|\cos(\theta_i - \frac{\pi}{2})|,\\
SqRel &= \frac{1}{|N|}\sum_{i\in N}\frac{\|d_i - d_i^*\|^2}{d_i^*}\cdot|\cos(\theta_i - \frac{\pi}{2})|,\\
RMSE &= \sqrt{\frac{1}{|N|}\sum_{i\in N}\|d_i - d_i^*\|^2\cdot|\cos(\theta_i - \frac{\pi}{2})|},\\
RMSELog &= \sqrt{\frac{1}{|N|}\sum_{i\in N}\|\log d_i - \log d_i^*\|^2\cdot|\cos(\theta_i - \frac{\pi}{2})|}
\end{aligned}
\tag{12}
$$

where d_i and d_i^* denote the predicted and ground-truth depth at pixel i, θ_i is the colatitude of that pixel, and N is the number of valid pixels. Accuracy is computed as:

$$
Accuracy(\delta) = \frac{1}{N'}\sum_{i\in N'}\mathbb{I}(\max(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}) < \delta) \tag{13}
$$

where $\mathbb{I}(\cdot)$ is the indicator function. The thresholds are set to $\delta_1 = 1.25, \delta_2 = 1.25^2, \delta_3 = 1.25^3$.

C.3. Implementation Details

We build our SO(3)-Equivariant ViT-Adapter on 3 different sizes of ViT backbones, including ViT-S, ViT-B, and ViT-L, following the embedding dimensions used in Depth Anything V2 [19].

ESPM Configurations. Our Equivariant Spherical Prior Module (ESPM) consists of one S^2 convolution followed by four SO(3) convolutions. The input has 3 channels, and the output channels for the five layers are [20,32,64,128,256]. The corresponding maximum β of the convolution kernels are set to $[\frac{1}{32}\pi, \frac{1}{20}\pi, \frac{1}{16}\pi, \frac{1}{8}\pi, \frac{1}{6}\pi]$. The input bandwidth is initialized as $B = 64$ and gradually reduced across layers as [B, B, B/2, B/4, B/8]. For the last three layers, we apply a 1×1 convolution to project the feature dimension to match the ViT embedding size.

SO(3)-DCAM Configurations. For the SO(3) deformable cross-attention mechanism (SO(3)-DCAM), we set the number of feature levels to $S = 3$, consistent with the ESPM outputs. The number of attention heads for ViT-S/B/L is set to 6, 12, and 16, respectively. More heads correspond to denser sampling directions on the sphere around each query point. For each direction, we sample 4 spatial points, with the initial soft constraint for the farthest sampling distance set to $\beta_{max} = \frac{1}{8}\pi$. At each sampled location, 6 rotational samples are taken.

Feature Interaction Configurations. The S^2-SO(3) feature interaction blocks follow the configuration of the original ViT-Adapter [6], with the number of interaction blocks $N = 4$ and an FFN ratio of 0.25.

Input Resolution and Processing. The input images are resized to 518×1036 during training to match the 14×14 patch size. For evaluation, predictions are resized back to 512×1024 before computing metrics.

Training Setup. We train the model on the PNVS-hard dataset for 30 epochs using the AdamW optimizer with a batch size of 4, an initial learning rate of 1e-4, and a weight decay of 0.05. Unless otherwise specified, no data augmentation is applied, as we aim to evaluate the sim-to-real generalization enabled by rotation-equivariant modeling rather than data-driven regularization.

D. Full Module Effectiveness Ablation Results

This section provides the full quantitative and qualitative results of our module ablation study on the Matterport3D test split. Due to space limitations, we include only the key metrics in the main paper. Here, we report the complete results across all depth estimation metrics and provide visual comparisons.

D.1. Quantitative Results

Table 3 reports the complete quantitative ablation results on the Matterport3D test split, extending the subset presented in the main paper. Replacing the spatial prior module in ViT-Adapter with our ESPM improves δ_1 from 87.2% to 87.5%, indicating that rotation-equivariant local features are important for improving dense predictions performance on spherical data. Introducing SO(3) Sampling further raises δ_1 to 87.9% and reduces RMSE from 0.431 to 0.426, showing that SO(3) sparse sampling is better suited for retrieving relevant features on SO(3) group for spherical tokens. Finally, incorporating the SCRPE module decreases RMSE to 0.416 and increases δ_1 to 88.2%. This suggests that modeling the relative 3D rotational relationship between spherical queries and their corresponding SO(3) features is effective for measuring complementary correlations among features. In summary, the three modules proposed in the paper for extracting and preserving rotation-equivariance based on the ViT-Adapterare all essential. Together, they reduce Abs Rel from 0.093 to 0.089, RMSE from 0.431 to 0.416, and improve δ_1 from 87.2% to 88.2%.

D.2. Qualitative Comparisons

Fig. 1 provides additional qualitative comparisons corresponding to the quantitative results above. The improvements brought by each module are clearly observable in regions with strong projection distortion, particularly at high latitudes. For example, the chandelier in the upper part of the

Method	ESPM	SO(3) Sampling	SCRPE	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE(log_{10}) ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
ViT-S(*) [19]				0.100	0.082	0.447	0.068	0.858	0.965	0.988
ViT-S [19]				0.093	0.076	0.439	0.065	0.871	0.965	0.987
ViT-Adapter [6]				0.093	0.071	0.431	0.065	0.872	0.966	0.987
Ours w/o SO(3)-DCAM	✓			0.092	0.071	0.430	0.065	0.875	0.966	0.987
Ours w/o SCRPE	✓	✓		0.091	0.071	0.426	0.064	0.879	0.968	0.988
Ours	✓	✓	✓	**0.089**	**0.069**	**0.416**	**0.063**	**0.882**	**0.969**	**0.988**

Table 3. Complete ablation results of main components on the Matterport3D test split. Values extend the subset reported in the main paper by including all error and accuracy metrics.

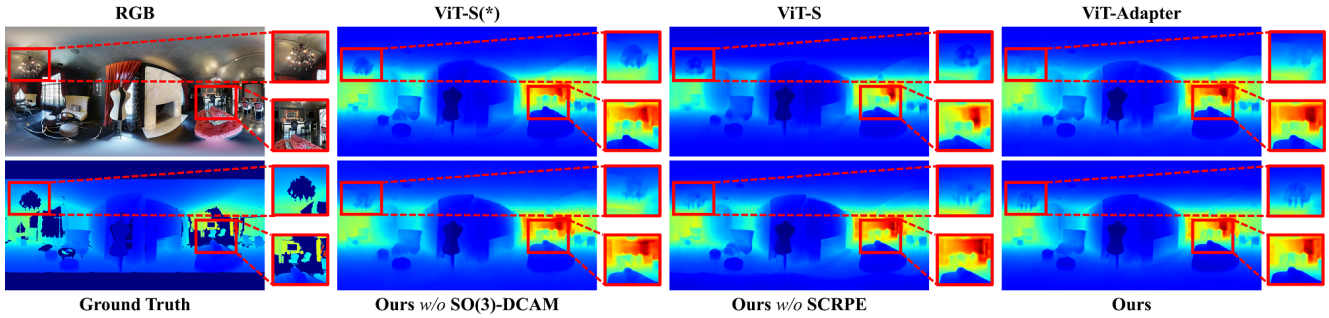

Figure 1. Additional qualitative comparisons within our module ablation study framework.

panorama is reconstructed with sharper structure and more accurate depth in our full model than in any of the baselines or partial variants. This visual evidence aligns with the quantitative findings that ESPM introduces cleaner, rotation-equivariant local features, and SO(3)-DCAM retrieves more geometrically consistent features on the rotation group. On the right side of the scene, only ViT-S(*) and our full model correctly distinguish the bar stools from the distant counter, demonstrating that our module design introduces minimal noise into the backbone and preserves the pre-trained depth priors learned by the ViT. Overall, the qualitative comparisons show that all proposed modules contribute meaningfully to depth accuracy and structural fidelity, validating the design of the full SO(3)-Equivariant ViT-Adapter.

E. Rotation Consistency under Horizontal Rotations

We further evaluate the rotation consistency of depth predictions under rotations around the gravity axis. Given an input panorama, we first obtain a reference depth prediction from the original image. We then horizontally translate the image by different pixel offsets and feed the shifted images into the network. The translation offsets range from 0 to $W - 14$, with a step size of 14 pixels, where $W = 1036$ denotes the image width. The step size corresponds to a one-token shift in the ViT backbone. For each offset, the predicted depth map is circularly shifted back to align with the original image, and we compute the L_2 difference between the restored depth and the reference prediction. The difference is measured in meters and averaged over valid pixels and all images in the Stanford2D3D area 5 test set. The results are shown in Fig. 2. Both methods exhibit very small prediction differences under horizontal translations (approximately 1–2 cm on average), while our method consistently produces slightly lower errors than the ViT-Adapter baseline. The small prediction differences are partly due to horizontal rotations preserving the latitude of scene content and therefore not introducing additional complex structures into high-latitude regions.

F. More Qualitative

We demonstrate additional zero-shot sim-to-real qualitative results in Fig. 3 and Fig. 4. Fig. 3 shows the results on the Matterport3D dataset, where the maximum value of the error map is set to 3m. Fig. 4 shows the results on the Stanford2D3D dataset, where the maximum value of the error map is set to 2m.

G. Limitation and Future Work

Although our method enhances rotation awareness through SO(3)-equivariant feature injection, the backbone itself remains inherently Euclidean. This architectural mismatch limits the model's ability to maintain globally consistent predictions under large 3D rotations. In practice, our framework handles small vertical perturbations (e.g., $\leq$30°) well during zero-shot inference, but extreme rotations (e.g., $\geq$90°) can still lead to structural distortions and degraded point cloud reconstruction.

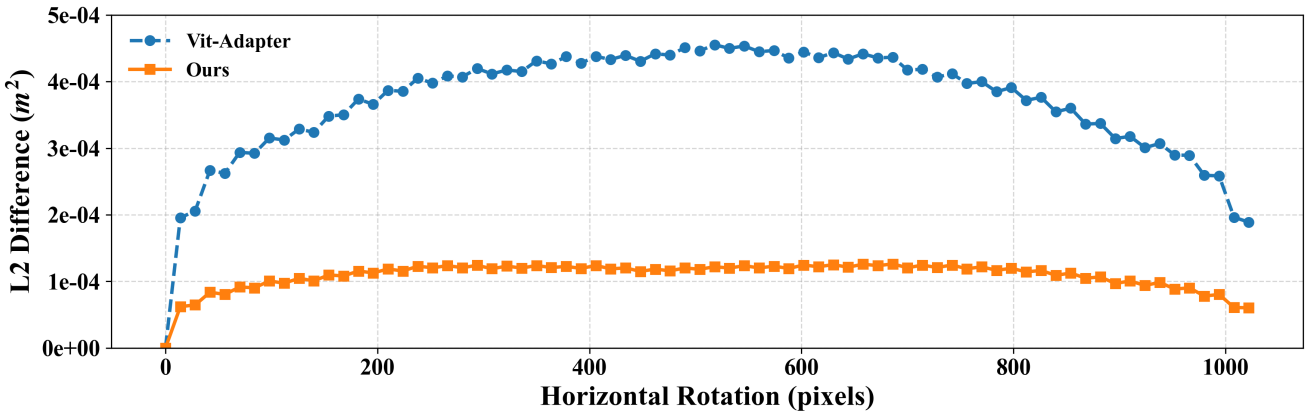

Figure 2. Rotation consistency under horizontal rotations. Lower values indicate stronger rotation consistency.

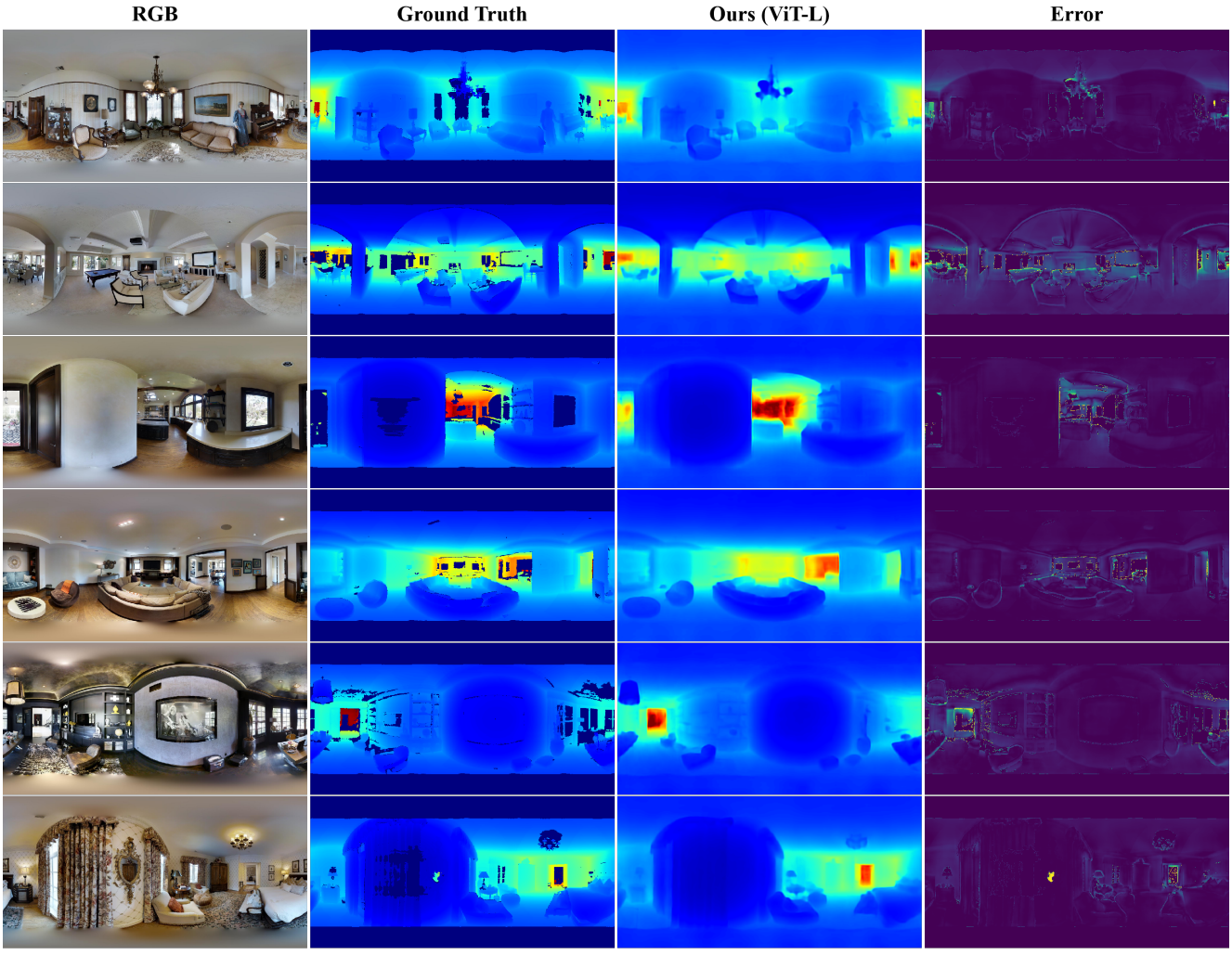

Figure 3. Additional qualitative results on the Matterport3D dataset with zero-shot sim-to-real setting.

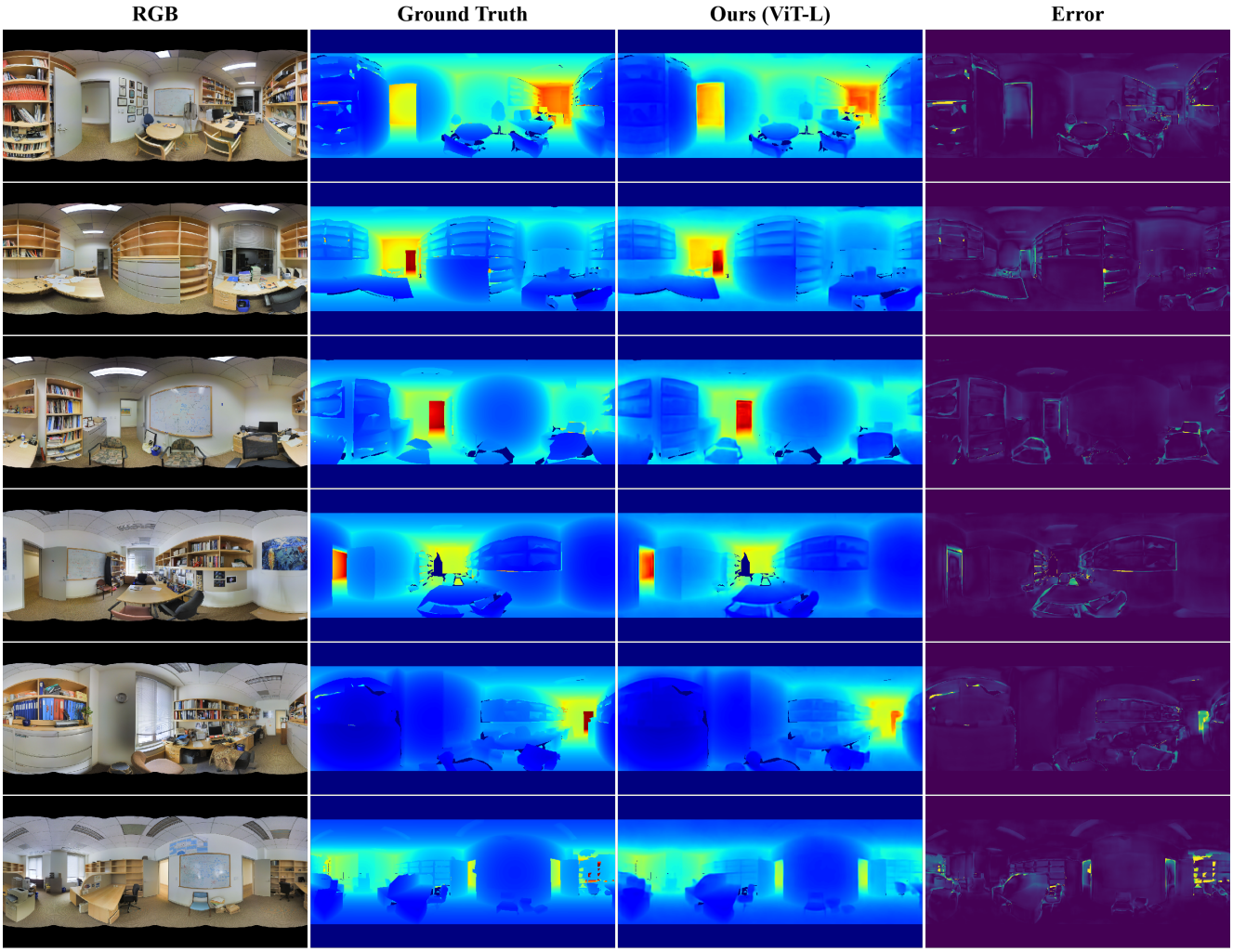

Figure 4. Additional qualitative results on the Stanford2D3D dataset with zero-shot sim-to-real setting.

Future research may move beyond feature-level alignment and investigate whether Euclidean pre-trained transformers can be progressively transformed into operators that behave equivariantly on spherical domains. Instead of only modifying inputs or injecting local equivariant priors, possible directions in future work include reparameterizing attention layers through representation and transforming pre-trained weights across Euclidean and spherical manifolds. Our further goal is to explore principled ways of transferring large scale perspective pretrained ViTs into models whose internal computations are more naturally compatible with SO(3) symmetries without relying on large panoramic training datasets.

References

[1] Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, and Lin Wang. HRDFuse: Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13273–13282, 2023. 3

[2] Hao Ai and Lin Wang. Elite360D: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9926–9935, 2024. 3

[3] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2D-3D-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, pages 1–9, 2017. 4

[4] Zidong Cao, Jinjing Zhu, Weiming Zhang, Hao Ai, Haotian Bai, Hengshuang Zhao, and Lin Wang. PanDA: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–16, 2025. 3, 4

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy

Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *IEEE International Conference on 3D Vision*, pages 667–676, 2017. 4

[6] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *International Conference on Learning Representations*, pages 1–20, 2023. 1, 5, 6

[7] Taco S. Cohen, Mario Geiger, Jonas Koehler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, pages 1–15, 2018. 2

[8] Dongyoon Hwang, Byungkun Lee, Hojoon Lee, Hyunseung Kim, and Jaegul Choo. Adapting pretrained ViTs with convolution injector for visuo-motor control. In *International Conference on Machine Learning*, volume 235, pages 20871–20888, 2024. 1

[9] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. UniFuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters*, 6(2):1519–1526, 2021. 3

[10] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. OmniFusion: 360 monocular depth estimation via geometry-aware fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2791–2800, 2022. 3

[11] E. B. Saff and A. B. J. Kuijlaars. Distributing many points on a sphere. *The Mathematical Intelligencer*, 19(1):5–11, 1997. 5

[12] 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, 2022. 3

[13] Cheng Sun, Min Sun, and Hwann-Tzong Chen. HoHoNet: 360 indoor holistic understanding with latent horizontal features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2573–2582, 2021. 3

[14] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2020. 3

[15] 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. 3

[16] Ning-Hsu Wang and Yu-Lun Liu. Depth Anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. In *Advances in Neural Information Processing Systems*, volume 37, pages 127739–127764, 2024. 3, 4

[17] Jiale Xu, Jia Zheng, Yanyu Xu, Rui Tang, and Shenghua Gao. Layout-guided novel view synthesis from a single indoor panorama. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16433–16442, 2021. 4

[18] Qingsong Yan, Qiang Wang, Kaiyong Zhao, Jie Chen, Bo Li, Xiaoweo Chu, and Fei Deng. SphereFusion: Efficient panorama depth estimation via gated fusion. In *International Conference on 3D Vision*, pages 855–865, 2025. 3

[19] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, pages 21875–21911, 2024. 5, 6

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

[21] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large photo-realistic dataset for structured 3D modeling. In *European Conference on Computer Vision*, page 519–535, 2020. 4

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