Elite360D: Towards Efficient 360 Depth Estimation via Semantic- and Distance-Aware Bi-Projection Fusion

Hao Ai[1] Lin Wang[1,2*]
[1]AI Thrust, HKUST(GZ) [2]Dept. of CSE, HKUST
hai033@connect.hkust-gz.edu.cn, linwang@ust.hk

Abstract

360 depth estimation has recently received great attention for 3D reconstruction owing to its omnidirectional field of view (FoV). Recent approaches are predominantly focused on cross-projection fusion with geometry-based re-projection: they fuse 360 images with equirectangular projection (ERP) and another projection type, e.g., *cubemap projection to estimate depth with the ERP format. However, these methods suffer from 1) limited local receptive fields, making it hardly possible to capture large FoV scenes, and 2) prohibitive computational cost, caused by the complex cross-projection fusion module design. In this paper, we propose* ***Elite360D****, a novel framework that inputs the ERP image and icosahedron projection (ICOSAP) point set, which is undistorted and spatially continuous. Elite360D is superior in its capacity in learning a representation from a local-with-global perspective. With a flexible ERP image encoder, it includes an ICOSAP point encoder, and a Bi-projection Bi-attention Fusion (B2F) module (totally ∼1M parameters). Specifically, the ERP image encoder can take various perspective image-trained backbones (*e.g., *ResNet, Transformer) to extract local features. The point encoder extracts the global features from the ICOSAP. Then, the B2F module captures the semantic- and distance-aware dependencies between each pixel of the ERP feature and the entire ICOSAP feature set. Without specific backbone design and obvious computational cost increase, Elite360D outperforms the prior arts on several benchmark datasets.*

Multimedia Material

For videos, code, demo and more information, you can visit https://VLIS2022.github.io/Elite360D/

1. Introduction

360° images capture the complete surrounding environment in one shot with an ultra-wide field-of-view (FoV) of $180^\circ \times 360^\circ$, which is broadly applied to applications, *e.g.*, autonomous driving [1, 19, 42, 43, 49] and virtual reality

*Corresponding author (e-mail: linwang@ust.hk)

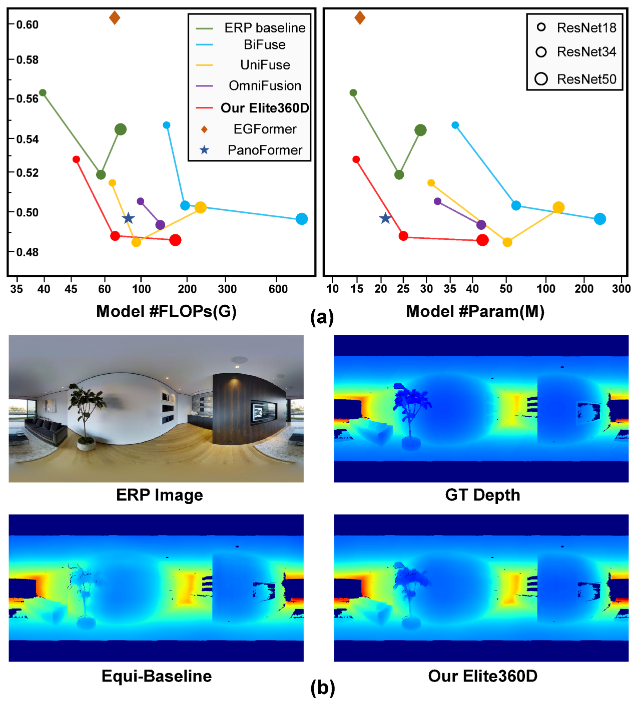

Figure 1. (a) Performance (RMSE error) curves on M3D test dataset [5]. Larger circles mean more parameters (*e.g.*, ResNet18, ResNet34, ResNet50) and lower errors mean better performance. (b) Comparison with the ResNet34 as the ERP encoder backbone. With only 1M more parameters, our depth result is more accurate.

[4, 23, 26]. The ability to infer the 3D structure of a surrounding scene has inspired active research for monocular 360 depth estimation. Equirectangular projection (ERP) is the commonly used projection type that can provide a complete view of a scene. An ERP image, *a.k.a.*, panorama, samples pixels with a higher density at the poles compared to the equator, resulting in spherical distortions (Fig. 2(c)).

In recent years, to address the spherical distortions, several ah-hoc designs [33, 36, 50, 51] focus on the distortion-aware convolutional kernels and sample grids from undistorted tangent planes on the sphere. These grids are back-projected to the corresponding locations of the ERP image. Also, some methods [28, 34, 44] partition the ERP input into multiple vertical slices and extract a compact representation from these slices to extend the receptive field of each pixel for the ERP representation. Unfortunately, these methods result in substantial computational costs. Meanwhile, the limited local receptive field of the convolutional filters is insufficient to provide the global perception of the panorama. Therefore, PanoFormer [32] and EGFormer [45] design specific transformer architectures to effectively model the long-range dependencies between the global context of large-FoV ERP images. Nevertheless, the receptive fields are confined by the sizes of local attention windows. Furthermore, the aforementioned data-driven methods often yield sub-optimal results when trained from scratch on limited 360° depth dataset (See Tab. 2).

With the power of existing large-scale perspective images [11], several methods [2, 16, 37, 38] employ the pre-trained models – designed for perspective images – as encoder backbones to extract ERP feature maps, and propose the cross-projection fusion to rectify distortions in the ERP feature maps. Specifically, these methods introduce the less-distorted planar projection data, *i.e*., cubemap projection (CP) patches or tangent projection (TP) patches (See Fig. 2(d), (e)), and unify the spatial dimensions between different projections for the cross-projection fusion. BiFuse and UniFuse [16, 37, 38] propose the C2E module to re-project the content in patch-wise CP features into ERP grid following the spherical geometric relationships. Furthermore, they leverage the fusion between ERP feature maps and C2E feature maps to enhance depth estimation accuracy. However, their concatenation-based feature fusion across multiple scales brings a significant computational burden. Based on the feature similarity, HRDFuse [2] spatially aligns the ERP pixel features and TP patch features. This makes it possible to significantly marry the global context in the ERP feature map and regional structural details in the TP patch features. However, the decoder-level fusion remarkably increases the computational memory and cost.

These less-distorted gnomonic projections, *i.e*., CP and TP patches, are spatially discontinuous and require complex re-projection operations for the desired ERP format predictions. Inspired by these issues, we introduce a more powerful projection, the icosahedron projection (ICOSAP) [21, 46], see Fig. 2(f). Importantly, ICOSAP is a spatially continuous and globally perceptive non-Euclidean projection for 360° images. In light of this, we propose an efficient and effective fusion-based framework, named **Elite360D** that takes the best of ERP and ICOSAP by learning a representation from a local-with-global perspective. It comprises three components: an ERP encoder, an ICOSAP encoder, and the bi-projection bi-attention fusion (B2F) module. For the ERP image encoder, our Elite360D flexibly supports a wide range of off-the-shelf 2D models, *e.g*., ResNet [14], Swin transformer [24], as the backbone to extract ERP feature maps. This potentially reduces the overfitting problems, particularly on small-scale 360 depth estimation datasets [3, 5]. For efficiency, we represent the ICOSAP spheres as the discrete point sets rather than icosahedron meshes [10, 15, 46] or unfolded representations [21, 43]. As such, we can avoid semantic information redundancy due to dense ERP pixels and maintain the spatial position information. Then, to enable each ERP pixel feature with local receptive fields to perceive the whole scene, B2F module captures the semantic- and distance-aware dependencies between each ERP pixel feature and entire ICOSAP feature set.

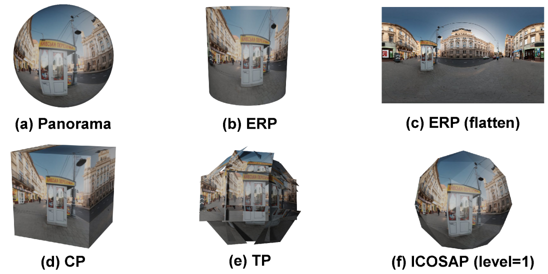

Figure 2. Different projections of a spherical imaging panorama.

We conduct extensive experiments on three different datasets with different encoder backbones to demonstrate the flexibility and effectiveness of our *Elite360D*. The experimental results indicate that our Elite360D significantly improves plain-backbones' performance with minimal computational memory (*only about 1M parameters*) (See Fig. 1(a)). Note that, based on the simple ERP depth estimation baseline, *Elite360D* only performs bi-projection feature fusion at the last feature layer and achieves results that are on par with leading methods, such as [2, 41].

In summary, our main contributions are three-fold: (I) We introduce the ICOSAP with spatial continuity and global perception and represent them as the discrete point sets to reduce the computation cost and maintain the spatial information; (II) We propose the B2F module that jointly models the semantic and spatial dependencies between ICOSAP and ERP features and learn the representations with global-with-local receptive fields for large-FoV scenes; (III) Building upon the B2F module, we propose a novel framework that supports diverse off-the-shelf models as encoder backbones, showing better flexibility than the specially designed models, *e.g*., [32, 45].

2. Related Works

Monocular 360 depth estimation Existing methods can be categorized into two types: one with single projection input, and the other with bi-projection inputs.

1) Single Projection Input With the ERP images as the inputs, many methods focus on mitigating the inherent spherical distortion issues. Several methods [8, 36, 51] rectify distortions by modifying the receptive fields of tradi-

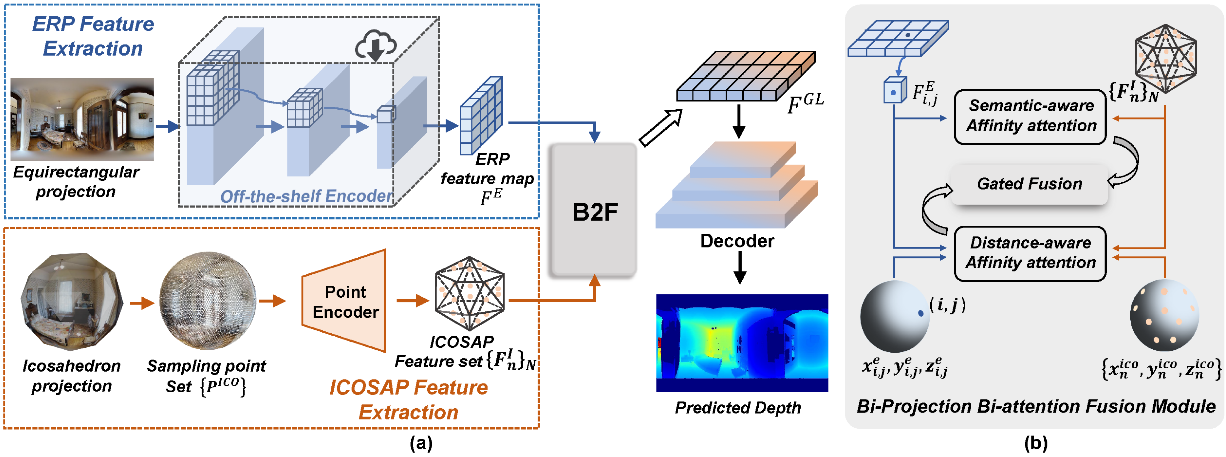

Figure 3. (a) **An overview of our *Elite360D framework***, comprising image-based ERP feature extraction (Sec. 3.1), point-based ICOSAP feature extraction (Sec. 3.2), and Bi-projection Bi-attention fusion (B2F) (Sec. 3.3). For better visualization, we do not show the skip connections [30] at the decoding stage. (b) Illustration of the B2F module, consisting of three parts: semantic-aware affinity attention block (Fig. 5), distance-aware affinity attention block (Fig. 6) and gated fusion (Eq. 5).

tional convolutional kernels. Besides, OmniFusion [41] and 360MonoDepth [29] follow [13] to predict depth maps from less distorted TP patches. Similarly, PanelNet [44] predicts depth maps on the vertical slices of ERP images and then aggregates them. Moreover, some methods [28, 34, 50] focus on tackling the small receptive fields of convolution filters for processing large FoV scenes. More recently, PanoFormer [32] and EGFormer [45] design transformer architectures [12] for ERP images to model the long-range dependencies. By contrast, S^{2}Net [22] extracts robust features from ERP images with the Swin transformer [24] and projects the ERP feature maps onto the spherical surface to minimize distortions and maintain spatial consistency.

2) Bi-Projection Inputs. BiFuse [37] utilizes a dense fusion strategy to bidirectionally fuse ERP and CP features at both encoding and decoding stages. In contrast, BiFuse++[38] and UniFuse [16] fuse ERP and CP features at the encoding stage. Recently, HRDFuse [2] achieved the SOTA performance based on the adaptive fusion of ERP and TP predictions. By contrast, we introduce a non-Euclidean projection, ICOSAP, which is less distorted and spatially continuous. We leverage the global perception capacity of ICOSAP to enable each ERP pixel-wise feature, with a limited local receptive field, to capture the entire scene.

Representations for 360 Images ERP is the most commonly used projection, which projects a panorama onto a cylinder (Fig. 2(b)) and then unfolds it into a plane (Fig. 2(c)). However, ERP maps latitudes and longitudes onto the vertical and horizontal axes with equal spacing on the plane, leading to distortions. CP [27] projects the panoramas onto six faces of a cube where each face is less distorted (Fig. 2(d)). Building upon CP, some padding methods [7, 37] remove the boundary inconsistency between the adjacent CP patches. Recently, Eder *et al.* [13] proposed the tangent projection (TP) (Fig. 2(e)), to significantly mitigate spherical distortion. Besides, [21, 31, 43] employ the polyhedrons, *e.g.*, octahedron and icosahedron, to represent the panoramas and unfold these polyhedron representations on the plane for processing. In Elite360D, we select popular ICOSAP (See Fig. 4) as our input. Notably, we represent ICOSAP as discrete points, significantly reducing the computational cost while preserving spatial information and global awareness. Meanwhile, with the dense pixels in the ERP panoramas, discrete ICOSAP points can avoid semantic information redundancy.

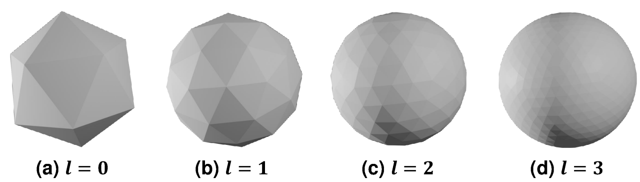

Figure 4. The subdivision of icosahedron at different resolution l.

Cross-Attention Mechanism It is widely used for efficient multi-modal feature fusion. For instance, Chen *et al.* [6] built cross-attention blocks to align and fuse 2D image features and 3D point cloud features for 3D object recognition. Similarly, BEVGuide [25], using BEV embedding as a guided query, employs a cross-attention block to fuse information across different sensors. We propose the B2F module to model the relationships between each pixel of the ERP feature and the whole ICOSAP point feature set, based on the semantic-aware affinity attention and distance-aware affinity attention simultaneously. With the B2F module, each ERP pixel feature can perceive the spatial information and semantic information of the large FoV scenes.

3. The Proposed Framework

Overview. Our goal is to investigate an efficient and effective bi-projection fusion framework for 360 depth estimation. To this end, as illustrated in Fig. 3, the proposed

Elite360D consists of three key components: image-based ERP feature extraction, point-based ICOSAP feature extraction and bi-projection bi-attention fusion (B2F). For the ERP feature F^E extraction, Elite360D flexibly accommodates a wide range of 2D models as encoder backbones and takes advantage of pre-training to reduce the over-fitting problems. For the ICOSAP feature extraction, we represent ICOSAP sphere as the point set, which significantly reduces the computation cost and preserves the spatial information and global perception. Then we directly employ the simple point encoder, *i.e*. Point transformer [47], to extract the ICOSAP point feature set $\{F_n^I\}_{n=1,\cdots,N}$, N is the point number, denoted as $\{\boldsymbol{F_n^I}\}_{\boldsymbol{N}}$. Next, to facilitate bi-projection feature fusion and learn a powerful representation F^{GL} from a local-with-global perspective, B2F module comprises semantic-aware affinity attention block, distance-aware affinity attention block, and gated fusion block. These two attention blocks model the semantic- and distance-aware dependencies between each ERP pixel-wise feature and entire ICOSAP feature set, respectively and the gated fusion block adaptively fuses the bi-attention outputs. Lastly, we employ a simple decoder with up-sampling and skip-connections to predict the final depth map from the fused representation F^{GL}. We now describe the details.

3.1. ERP Feature Extraction

Taking an ERP image with the resolution size of $H \times W$ as the input, the encoder extracts feature map $F^E \in \mathbb{R}^{h\times w\times C}$, where $h = H/s$, $w = W/s$, s is the down-sampling scale factor and C is the channel number. In particular, as treating ERP images as 2D perspective images, our encoder backbone is compatible with a wide range of robust 2D models that have been pre-trained on large-scale perspective image datasets [11], including CNNs and vision transformers, *e.g*., ResNet [14], EfficientNet [35], Swin transformer [24]. Notably, as repeated local operations cannot substitute for a global operator [40, 45], the finite size of convolutional kernels or the limited size of local attention windows results in ERP pixel-wise features lacking sufficient global receptive fields. Additionally, these 2D model performance suffers from the spherical distortions (See Tab. 1).

3.2. ICOSAP Feature Extraction

Preliminary. Since Lee *et al*. [21] found that polyhedron representations with a greater number of initial faces exhibit lower distortion, this paper focuses on the ICOSAP with the most initial faces. Especially, we introduce the ICOSAP to offer a comprehensive and high-quality global perception. For the resolution, ICOSAP data consists of 20×4^l faces and 12×4^l vertices at the subdivision level l (See Fig. 4). Given the spatial relationships between ICOSAP and ERP within the spherical space, RGB values of ICOSAP vertices can be derived from corresponding ERP pixels.

Feature extraction. Existing works focus on processing

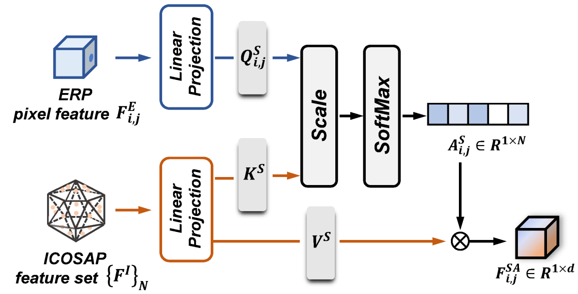

Figure 5. The architecture of semantic-aware affinity attention. Especially, $\mathbf{Q}^S_{i,j} \in \mathbb{R}^{1\times d}$, $\mathbf{K}^S \in \mathbb{R}^{N\times d}$, and $\mathbf{V}^S \in \mathbb{R}^{N\times d}$, where d is the dimension and N is the ICOSAP point number.

ICOSAP data on the plane, *i.e*., unfolded mesh representation [46] and spherical polyhedron representations [21]. However, these methods require specially designed and computationally intensive operations, such as convolutions, pooling [21], and up-sampling [31], to process the planar representations. By contrast, we propose to represent an ICOSAP input as a point set, as illustrated at the bottom of Fig. 3(a). ERP images provide dense pixel values and discrete ICOSAP point set can prevent redundancy in semantic information while preserving spatial information and global perception. Therefore, we take full advantage of two projections for efficient and accurate 360 depth estimation. In detail, we employ the central points of each face to represent the ICOSAP sphere. Firstly, we obtain the 20×4^l faces of an ICOSAP sphere at the default subdivision level l, where each face is composed of three vertices. Then, we calculate the spatial coordinates and RGB values of each face center by averaging them of each three vertices, as each face is an equilateral triangle. As a result, we obtain the ICOSCAP point set $\{\boldsymbol{P^{ICO}}\} \in \mathbb{R}^{(20\times 4^l)\times 6}$, where 20×4^l is the point number and 6 represents the coordinates $[x, y, z]$ and RGB channels. With the input point set $\{\boldsymbol{P^{ICO}}\}$, we directly employ the encoder of Point Transformer [47] to extract ICOSAP point feature set $\{\boldsymbol{F_n^I}\}_{\boldsymbol{N}} \in \mathbb{R}^{N\times C}$, where N represents the number of point features and C corresponds to the same channel number as the ERP feature map F^E. Especially, for simplicity and efficiency, we do not opt for complex networks for this purpose.

3.3. Bi-Projection Bi-Attention Fusion

In earlier methods [37] and [16], bi-projection feature fusion primarily depends on the geometric relationships between CP and ERP. They first apply the C2E function to re-project CP feature patches into ERP format feature maps and subsequently perform pixel-wise feature concatenation. While this fusion is effective, it brings several problems: 1) geometry-based re-projection and concatenation-based feature fusion significantly increase the computational costs (See Tab. 2); 2) the geometry-based fusion process from CP patches to ERP panorama restricts ERP pixels from the global perception, where each ERP pixel perceives scene in-

Datasets	Backbone	Method	#Params (M)	#FLOPs (G)	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$\delta_1(\%)\uparrow$	$\delta_2(\%)\uparrow$	$\delta_3(\%)\uparrow$
M3D [5]	ResNet-18 [14]	Equi-Baseline	14.39	39.84	0.1428	0.1203	0.5607	80.97	94.68	98.26
		Ours	15.43	45.91	0.1272	0.1070	0.5270	85.28	95.28	98.49
		Δ	+1.04	+6.07	**-11.06%**	**-9.73%**	**-6.01%**	**+4.31**	**+0.60**	**+0.23**
	ResNet-34 [14]	Equi-Baseline	24.50	59.27	0.1255	0.1048	0.5173	85.52	96.10	98.55
		Ours	25.54	65.29	0.1115	0.0914	0.4875	88.15	96.46	98.74
		Δ	+1.04	+6.02	**-11.16%**	**-12.79%**	**-5.76%**	**+2.63**	**+0.36**	**+0.19**
	ResNet-50* [14]	Equi-Baseline	29.63	75.13	0.1370	0.1209	0.5432	82.51	94.83	98.07
		Ours	42.99	170.11	0.1112	0.0980	0.4870	86.70	96.01	98.61
		Δ	+13.36	+94.98	**-18.75%**	**-18.94%**	**-10.35%**	**+4.19**	**+1.18**	**+0.54**
	EfficientNet-B5 [35]	Equi-Baseline	33.10	18.81	0.1034	0.0831	0.4638	89.92	96.80	98.80
		Ours	34.11	24.88	0.1048	0.0805	0.4524	89.92	97.07	99.14
		Δ	+1.01	+6.07	**+1.35%**	**-3.13%**	**-2.46%**	-0.00	**+0.27**	**+0.34**
	SWin-B* [24]	Equi-Baseline	90.75	187.37	0.1301	0.1250	0.5557	84.06	94.76	97.72
		Ours	94.30	211.28	0.1249	0.1210	0.5462	85.34	95.07	97.77
		Δ	+3.55	+23.91	**-4.00%**	**-3.20%**	**-1.71%**	**+1.28**	**+0.31**	**+0.05**
	DilateFormer-T [17]	Equi-Baseline	19.93	46.96	0.1429	0.1274	0.5748	81.15	94.46	97.79
		Ours	20.94	53.02	0.1423	0.1251	0.5517	82.68	95.00	98.14
		Δ	+1.01	+6.06	**-0.42%**	**-1.81%**	**-4.02%**	**+1.53**	**+0.54**	**+0.35**
S2D3D [3]	ResNet-34 [14]	Equi-Baseline	24.50	59.22	0.1203	0.0754	0.3724	87.41	96.51	98.77
		Ours	25.51	65.28	0.1182	0.0728	0.3756	88.72	96.84	98.92
		Δ	+1.01	+6.06	**-9.21%**	**-3.45%**	+0.86%	**+1.31**	**+0.33**	**+0.15**
	EfficientNet-B5 [35]	Equi-Baseline	33.10	18.81	0.1026	0.0638	0.3580	89.43	97.06	99.16
		Ours	34.11	24.88	0.1018	0.0603	0.3575	89.47	97.23	99.22
		Δ	+1.01	+6.07	**-0.78%**	**-5.49%**	**-0.14%**	**+0.03**	**+0.17**	**+0.06**
Struct3D [48]	ResNet-34 [14]	Equi-Baseline	24.50	59.22	0.2256	0.3910	0.7641	76.90	89.96	94.49
		Ours	25.51	65.28	0.1480	0.2215	0.4961	87.41	94.34	96.66
		Δ	+1.01	+6.06	**-34.40%**	**-43.35%**	**-35.07%**	**+10.51**	**+4.38**	**+2.17**
	EfficientNet-B5 [35]	Equi-Baseline	33.10	18.81	0.1312	0.1938	0.4312	88.53	95.21	97.40
		Ours	34.11	24.88	0.1277	0.1930	0.4151	89.16	95.33	97.43
		Δ	+1.01	+6.07	**-2.67%**	**-0.41%**	**-3.73%**	**+0.63**	**+0.12**	**+0.03**

Table 1. **Quantitative comparison with ERP-based depth baseline**. For most conditions, we set the channel number C to 64. Especially, we set C to 256 for ResNet-50* and to 128 for Swin-B*. **Bold** indicates performance improvement. **Red** indicates performance decline.

formation only from its corresponding small-FoV CP patch without considering other patches; 3) this one-to-one alignment pattern only emphasizes spatial consistency, with no consideration for the semantic similarity. Thus, we design the Bi-Projection Bi-Attention Fusion (B2F) module to solve the above problems. As shown in Fig. 3(b), B2F module first leverages semantic-aware affinity attention and distance-aware affinity attention to model the semantic and spatial dependencies between each ERP pixel-wise feature $F^E_{i,j}$ and ICOSAP point feature set $\{\boldsymbol{F^I_n}\}_{\boldsymbol{N}}$. Here (i,j) indicates the coordinate of a pixel in the ERP feature map, $i \in (1,h), j \in (1,w)$, and N is the point number of the ICOSAP feature set. Consequently, a gated fusion block is employed to adaptively marry semantic- and distance-related information. **Semantic-aware affinity attention.** The semantic-aware affinity attention, as shown in Fig. 5, follows the the standard dot-product attention [12]. In detail, given the extracted ERP feature map $F^E \in \mathbb{R}^{h\times w\times C}$ and ICOSAP point feature set $\{\boldsymbol{F^I_n}\}_{\boldsymbol{N}} \in \mathbb{R}^{N\times C}$ (for simplicity, we denote it as $[\boldsymbol{F^I}]$), we generate the query $\mathbf{Q}^S_{i,j}$ from each ERP pixel-wise feature $F^E_{i,j} \in \mathbb{R}^{1\times C}$ and produce the key $\mathbf{K}^S$ and value $\mathbf{V}^S$ from the whole ICOSAP point feature set $[\boldsymbol{F^I}]$. Then, we calculate the attention weight $\mathbf{A}^S_{i,j}$ based on $\mathbf{Q}^S_{i,j}$ and $\mathbf{K}^S$, and obtain the affinity feature $F^{SA}_{i,j}$ with $F^{SA}_{i,j} = \mathbf{A}^S_{i,j} * \mathbf{V}^S, \mathbf{A}^S_{i,j} = softmax(\frac{\mathbf{Q}^S_{i,j}\mathbf{K}^{S^T}}{\sqrt{d}})$, where d is the channel number and

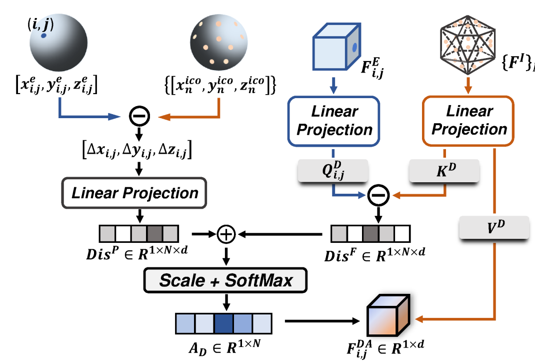

Figure 6. The architecture of distance-aware affinity attention, especially, $Q^D_{i,j} \in R^{1\times d}$, $K^D \in \mathbb{R}^{N\times d}$, and $\mathbf{V}^D \in \mathbb{R}^{N\times d}$.

$d = C$. After querying all the ERP pixel-wise features, we can obtain the ERP format feature map F^{SA} with the dimension of $\mathbb{R}^{h\times w\times d}$. The attention weight $A^S_{i,j}$ captures affinities based on the semantic-aware feature similarities, and the output F^{SA} effectively integrates the global and local receptive fields.

Distance-aware affinity attention. Distance-aware affinity attention captures the differences between ERP and ICOSAP in both spatial and semantic information. This can better utilize the geometric prior knowledge of panoramas and precisely measure the distances between two different projection features. As depicted in Fig. 6, the distance-aware affinity attention is built upon the subtraction-based

Datasets	Backbone	Method	Pub'Year	#Params (M)	#FLOPs (G)	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1(%) ↑	δ_2(%) ↑	δ_3(%) ↑
M3D [5]	Transformer	EGFormer [45]	ICCV'23	15.39	66.21	0.1473	0.1517	0.6025	81.58	93.90	97.35
		PanoFormer [32]	ECCV'22	20.38	81.09	0.1051	0.0966	0.4929	89.08	96.23	98.31
	ResNet-18 [14]	BiFuse [37]	CVPR'20	35.80	165.66	0.1360	0.1202	0.5488	83.27	95.12	98.10
		UniFuse [16]	RAL'21	30.26	62.60	0.1191	0.1030	0.5158	86.04	95.84	98.30
		OmniFusion [41]	CVPR'22	32.35	98.68	0.1209	0.1090	0.5055	86.58	95.81	98.36
		HRDFuse† [2]	CVPR'23	26.09	50.59	0.1414	0.1241	0.5507	81.48	94.89	98.20
		Ours	-	15.43	45.91	0.1272	0.1070	0.5270	85.28	95.28	98.49
	ResNet-34 [14]	BiFuse [37]	CVPR'20	56.01	199.58	0.1126	0.0992	0.5027	88.00	96.13	98.47
		BiFuse++ [38]	TPAMI'22	52.49	87.48	0.1123	0.0915	0.4853	88.12	96.56	98.69
		UniFuse [16]	RAL'21	50.48	96.52	0.1144	0.0936	0.4835	87.85	96.59	98.73
		OmniFusion [41]	CVPR'22	42.46	142.29	0.1161	0.1007	0.4931	87.72	96.15	98.44
		HRDFuse† [2]	CVPR'23	46.31	80.87	0.1172	0.0971	0.5025	86.74	96.17	98.49
		Ours	-	25.54	65.29	0.1115	0.0914	0.4875	88.15	96.46	98.74
	ResNet-50* [14]	BiFuse [37]	CVPR'20	253.08	775.24	0.1179	0.0981	0.4970	86.74	96.27	98.66
		UniFuse [16]	RAL'21	131.30	222.30	0.1185	0.0984	0.5024	86.66	96.18	98.50
		Ours	-	42.99	170.11	0.1112	0.0980	0.4870	86.70	96.01	98.61
S2D3D [3]	Transformer	EGFormer [45]	ICCV'23	15.39	66.21	0.1528	0.1408	0.4974	81.85	93.38	97.36
		PanoFormer [32]	ECCV'22	20.38	81.09	0.1122	0.0786	0.3945	88.74	95.84	98.59
	ResNet-34 [14]	OmniFusion [41]	CVPR'22	42.46	142.29	0.1154	0.0775	0.3809	86.74	96.03	98.71
		UniFuse [16]	RAL'21	50.48	96.52	0.1124	0.0709	0.3555	87.06	97.04	98.99
		Ours	-	25.51	65.28	0.1182	0.0728	0.3756	88.72	96.84	98.92
Struct3D [48]	Transformer	EGFormer [45]	ICCV'23	15.39	66.21	0.2205	0.4509	0.6841	79.79	90.71	94.55
		PanoFormer [32]	ECCV'22	20.38	81.09	0.2549	0.4949	0.7937	74.70	89.15	93.97
	ResNet-34 [14]	BiFuse [37]	CVPR'20	56.01	199.58	0.1573	0.2455	0.5213	85.91	94.00	96.72
		UniFuse [16]	RAL'21	50.48	96.52	0.1506	0.2319	0.5016	85.42	93.99	96.76
		Ours	-	25.51	65.28	0.1480	0.2215	0.4961	87.41	94.34	96.66

Table 2. **Quantitative comparison with the SOTA methods**. † means that we modify the HRDFuse network structure for a fair comparison. Green represents the best performance under the given encoder backbone.

cross-attention mechanism [39]. Given the ERP pixel-wise feature $F^E_{i,j}$ and ICOSAP point feature set $[\boldsymbol{F^I}]$, we firstly calculate the spatial distance embedding Dis^{SP} from the spatial coordinates of ERP pixel and ICOSAP point set, *i.e.* $[x^e_{i,j}, y^e_{i,j}, z^e_{i,j}]$ and $\{[x^{ico}_n, y^{ico}_n, z^{ico}_n]\}_N$, as:

$$Dis^{SP}_{i,j} = [e^{-\Delta x_{i,j}}, e^{-\Delta y_{i,j}}, e^{-\Delta z_{i,j}}]\mathbf{W}^{SP}, \qquad (1)$$

where linear projection $\mathbf{W}^{SP} \in \mathbb{R}^{3\times d}$, $Dis^{SP}_{i,j} \in \mathbb{R}^{1\times N\times d}$, and $[\Delta x_{i,j}, \Delta y_{i,j}, \Delta z_{i,j}] \in \mathbb{R}^{1\times N\times 3}$ is the distances between $[x^e_{i,j}, y^e_{i,j}, z^e_{i,j}]$ and $\{[x^{ico}_n, y^{ico}_n, z^{ico}_n]\}_N$. In particular, the operation $e^{-(\cdot)}$ allows Dis^{SP} to pay more attention to the close parts between ERP pixels and ICOSAP point set. After that, we produce the query $\mathbf{Q}^D_{i,j}$ and key $\mathbf{K}^D$ from $F^E_{i,j}$ and $[\boldsymbol{F^I}]$, respectively and calculate the semantic distance embedding $Dis^{SE}_{i,j}$:

$$Dis^{SE}_{i,j} = e^{-\left\|\mathbf{Q}^D_{i,j} - \mathbf{K}^D\right\|}, \qquad (2)$$

where $\mathbf{W}^D_Q$, $\mathbf{W}^D_K \in \mathbb{R}^{C\times d}$ are linear projections, $\mathbf{Q}^D_{i,j} \in \mathbb{R}^{1\times d}$, $\mathbf{K}^D \in \mathbb{R}^{N\times d}$, and $Dis^{SE}_{i,j} \in \mathbb{R}^{1\times N\times d}$. Lastly, the distance-aware attention weight $\mathbf{A}^D_{i,j}$ is generated with spatial and semantic distance embeddings, and the distance-aware affinity feature vector $F^{DA}_{i,j}$ is obtained from the attention weight $\mathbf{A}^D_{i,j}$ and the value $\mathbf{V}^D$:

$$\mathbf{A}^D_{i,j} = softmax(\frac{\sum(Dis^{SP}_{i,j} + Dis^{SE}_{i,j})}{\sqrt{d}}), \qquad (3)$$

$$\mathbf{V}^D = F^I\mathbf{W}^D_V, \;\; F^{DA}_{i,j} = \mathbf{A}^D_{i,j} * \mathbf{V}^D, \qquad (4)$$

where $\sum$ means the sum for the channel dimension, *i.e.* $\sum(Dis^{SP}_{i,j} + Dis^{SE}_{i,j}) \in \mathbb{R}^{1\times N}$. After querying all ERP pixel-wise features, we obtain the distance-aware aggregated feature $F^{DA} \in \mathbb{R}^{h\times w\times d}$, $d = C$.

Gated fusion. Since direct average or concatenation may compromise the original representation ability, inspired by [9], we propose the gated fusion block to adaptively fuse F^{SA} and F^{DA} and obtain the representations F^{GL} from a local-with-global perspective, formulated as:

$$\begin{aligned} F^{GL} &= g^{SA} * F^{SA} + g^{DA} * F^{DA}, \\ g^{SA} &= \sigma_{SA}(W^{SA}_g \cdot [F^{SA}; F^{DA}]), \\ g^{DA} &= \sigma_{DA}(W^{DA}_g \cdot [F^{SA}; F^{DA}]), \end{aligned} \qquad (5)$$

where W^{SA}_g and W^{DA}_g are linear projections, $\sigma_{SA}(\cdot)$ and $\sigma_{DA}(\cdot)$ are the sigmoid functions.

3.4. Optimization

With the fused feature F^{GL} and multi-scale ERP feature maps in the ERP encoder backbone, we feed them into a decoder [30] with several up-sampling blocks and skip-connections to output the final depth. For the depth supervision, we follow existing works [28, 32] and employ the combination of Berhu loss [20] and gradient loss [28]. (*Details of training loss can be found in the suppl. material*.).

4. Experiments

4.1. Datasets, Metrics, and Implementation Details

Datasets and Metrics. We evaluate Elite360D on three datasets: two real-world datasets, Matterport3D(M3D) [5], Stanford2D3D(S2D3D) [3], and a recently proposed large-scale synthetic dataset, Structure3D(Struct3D) [48]. For the evaluation metrics, we follow previous works [16, 32, 45] to employ some standard metrics, including absolute relative error (Abs Rel), squared relative error (Sq Rel), root mean squared error (RMSE), and three threshold percentage $\delta < \alpha^t$ ($\alpha = 1.25, t = 1, 2, 3$), denoted as δ_t. Addi-

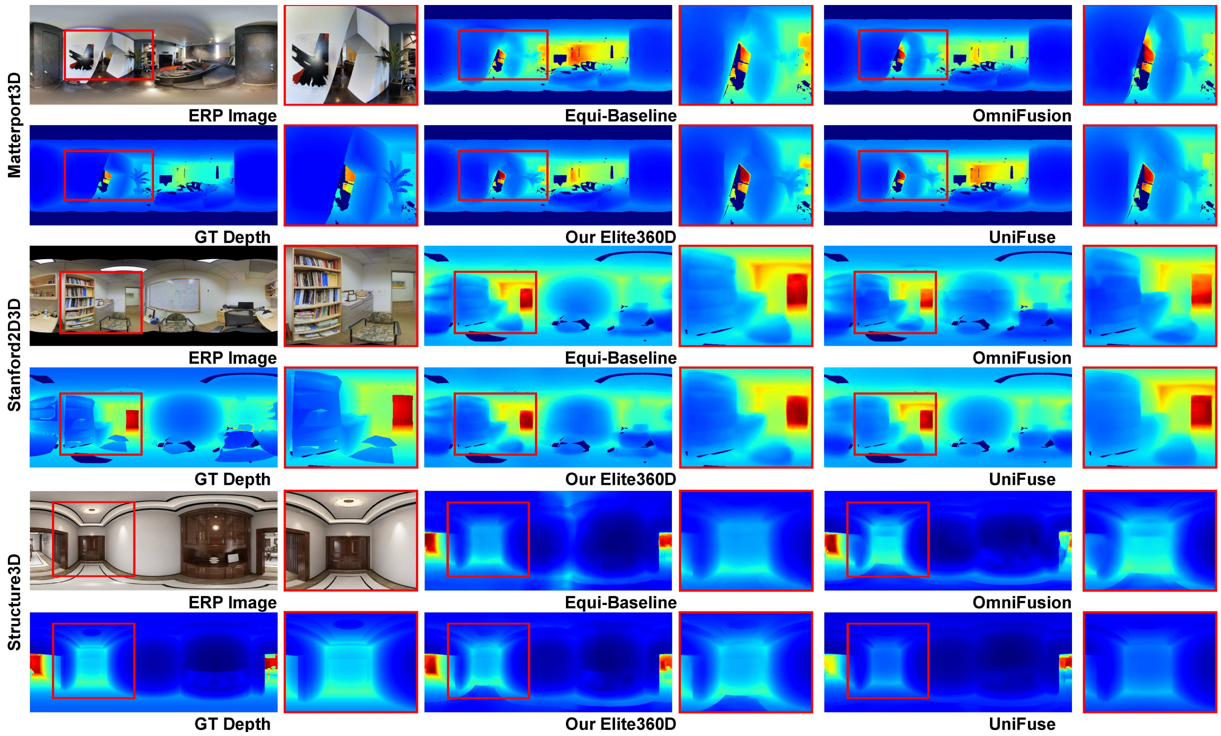

Figure 7. Qualitative results (with ResNet-34 as the backbone) on Matterport3D (top), Stanford2D3D (middle) and Structure3D (bottom).

Backbone	Pre-trained	Abs Rel↓	Sq Rel↓	RMSE↓	δ_1 ↑
ResNet-34	✗	0.1596	0.1452	0.5856	81.36
	✓	0.1115	0.0914	0.4875	88.15
EfficientNet-B5	✗	0.1211	0.1087	0.5131	86.80
	✓	0.1048	0.0805	0.4524	89.92
DilateFormer	✗	0.1515	0.1415	0.5694	80.95
	✓	0.1423	0.1251	0.5517	82.68

Table 3. The impact of pre-training on ERP encoder backbones.

tionally, we measure the number of parameters and FLOPS to evaluate the efficiency of our method.

Training Details. We use diverse ERP encoder backbones, including CNNs (ResNet-18, 34, 50 [14], EfficientNet B5 [35]), and transformers (Swin-B [24], DilateFormer-T [17]). All backbones are pre-trained on ImageNet-1K [11]. We set the default channel number C to 64 and default subdivision level of ICOSAP as $l = 4$. For the ICOSAP encoder, we employ the one of Point transformer [47] with three down-sample blocks. Following [16], we use Adam optimizer [18] and a constant learning rate of 1e^{-4}. Considering the unfair comparisons stemming from variations in hyper-parameters and validation procedures used across different methods, we re-train the existing methods from scratch and validate them, following the unified training and validation settings [16]. (*Due to page limit, detailed training and validation settings can be found in suppl. mat.*).

4.2. Performance Comparison

Comparisons with ERP-based depth baselines. As shown in Tab. 1, with an increase of only $\sim$1M parameters (C=64), our Elite360D demonstrates substantial advancements over the ERP-based baselines across different ERP encoder backbones on all three datasets. Specifically, for the Matterport3D dataset, Elite360D achieves reductions exceeding 10% in Abs Rel error (ResNet-18, 34), along with reductions of 4.00% in Abs Rel error (Swin-B) and 4.02% in RMSE error (DilateFormer-T). Besides, with the larger channel number $C = 256$ (ResNet-50), Elite360D outperforms ERP baseline by 18.75% (Abl Rel), 18.94% (Sq Rel). For the small-scale S2D3D dataset, Elite360D outperforms ERP baseline by 9.21% in Abs Rel error and 1.31% in accuracy δ_1 (ResNet-34), as well as 5.49% in Sq Rel error (EfficientNet-B5). Remarkably, on the larger-scale Structure3D, Elite360 performs favorably against the baseline by a significant margin, especially with ResNet-34.

Comparisons with prevalent methods. In Tab. 2, we conduct a comprehensive comparison with prevalent supervised methods. From the results, we can observe that our approach achieves similar or even superior performance compared to existing both bi-projection fusion methods and single input methods at a significantly lower cost, particularly on two large-scale datasets, Matterport3D and Structure3D. Specifically, for the Matterport3D dataset, our Elite360D with ResNet-34 outperforms UniFuse by 2.53% (Abs Rel) and with ResNet-50 outperforms BiFuse by 2.01% (RMSE). For the Structure3D dataset, our Elite360D with ResNet-34 outperforms UniFuse by 4.48% (Seq Rel), 1.99% (δ_1). For performance on the Stanford2D3D dataset, we suspect it might be related to the ICOSAP point encoder.

Bi-projection feature fusion	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1 ↑
SFA [2] + Add	0.1276	0.1002	0.5150	84.27
SFA [2] + Concat	0.1191	0.1019	0.5143	86.52
Only SA	0.1204	0.1014	0.5121	86.26
Only DA	0.1184	0.0972	0.4944	87.06
Our B2F (SA + DA)	**0.1115**	**0.0914**	**0.4875**	**88.15**

Table 4. The ablation results for B2F module.

Final fusion	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1 ↑
Add	0.1685	0.1481	0.5809	74.60
Average	0.1198	0.0918	0.4893	86.65
Concatenation	0.1145	0.0937	0.4880	87.66
Adaptive fusion [2]	0.1244	0.0968	0.4891	86.08
Our gated fusion	**0.1115**	**0.0914**	**0.4875**	**88.15**

Table 5. The ablation results for the fusion of B2F module.

The limited data of Stanford2D3D dataset restricts the ability of the transformer-based point encoder to provide accurate global perception. Moreover, in Fig. 7, we present the qualitative comparisons. Our Elite360D can predict more accurate depth values based on the local-with-global perception capabilities (e.g., flowers, shelves and doors). *Additional qualitative results and inference time comparisons can be found in the suppl. material.*

4.3. Ablation Study and Analyses

Most of ablation experiments are conducted on the Matteport3D test dataset with ResNet34 as the backbone .

The Effect of pre-training. We verify the effectiveness of ImageNet [11] pre-training with different encoder backbones. As observed from Tab. 3, the pre-training results in a significant improvement for all encoder backbones, *e.g.*, 6.79% improvement in accuracy δ_1 (ResNet-34). Notably, pre-training has a relatively small impact on DilateFormer. Combined with the results in Tab. 1, the explanation of this phenomenon is that the default input resolution in pre-trained models is different from actual input, thereby impacting the resolution-related position embeddings. In general, pre-training based on large-scale perspective images can effectively enhance the performance of models based on 360° images and reduce the risk of overfitting.

The effectiveness of B2F module. In Tab. 4, we compare four available bi-projection feature fusion modules. To align the spatial dimensions between ICOSAP point feature set and ERP feature map, we introduce SFA module from [2]. After that, we employ direct addition and concatenation to aggregate these two projections. We also achieve the bi-projection feature fusion with semantic-aware affinity attention (SA) alone and distance-aware affinity attention (DA) alone. Compared to the methods based solely on semantic-aware feature similarities (The first three rows), single distance-aware affinity attention can achieve better performance, which indicates that the spatial positional relationships boost the bi-projection feature fusion. Overall, our B2F module achieves the best performance.

The superiority of ICOSAP. As only CP/TP's patch centers lie on the sphere's surface, we extract the feature embedding from each CP/TP patch and employs the patch center coordinates and feature embedding as the input of B2F module. In Tab. 6, we show the results with ResNet18 backbone. Our Elite360D, utilizing the ICOSAP point set, marginally outperforms models with CP and TP patches, while exhibiting fewer parameters and FLOPs.

Method	#Param(M)	#FLOPs(G)	Abs Rel ↓	RMSE ↓	δ_1 ↑
ERP-CP	25.66	54.15	0.1369	0.5401	83.69
ERP-TP (N=18)	25.66	50.58	0.1328	0.5385	83.87
ERP-ICOSAP (Ours)	**15.43**	**45.91**	**0.1272**	**0.5270**	**85.28**

Table 6. The comparison of different projections on Matterport3D.

N of $\{F^I\}$	#Params (M)	#FLOPs (G)	Abs Rel↓	Sq Rel↓	RMSE↓	δ_1↑
20	27.41	66.29	0.1157	0.0995	0.5024	87.12
80	25.54	65.29	**0.1115**	**0.0914**	**0.4875**	**88.15**
320	24.98	64.32	0.1153	0.0943	0.4905	87.85

Table 7. Impact of the ICOSAP point-wise feature number N. Larger N, fewer down-sampling blocks in the point encoder.

The effectiveness of gated fusion. We conduct an ablation study for the gated fusion block, outlined in Tab. 5. With the feature maps F^{SA} and F^{DA}, We compare it with the direct addition, average fusion, concatenation, and the adaptive fusion in [2]. The gated fusion performs best.

ICOSAP point feature number N. We study the effect of the ICOSAP feature point number (See Tab. 7). Too few points (N=20) lead to the over-concentrated global contextual information resulting from excessive down-sampling blocks, while too many points (N=320) lead to under-concentrated condition, resulting in insufficient perception of ERP pixel features. Best performance can be observed when N=80 and we used N=80 as default in this paper.

5. Conclusion and Future Work

In this paper, we proposed a novel bi-projection fusion solution for efficient 360 depth estimation. To address the limited local receptive field of ERP pixel-wise features and avoid expensive bi-projection fusion modules, we proposed a compact yet effective B2F module to learn the representations with local-with-global perceptions from ERP and ICOSAP. With an increase of 1M parameters, we significantly improved the performance of the ERP-based depth estimation baseline. Remarkably, our approach achieved performance on par with complex state-of-the-art methods.

Future Work: From the experimental results, we observed that ERP-based depth baseline, with pre-trained EfficientNet backbone, even outperforms existing specifically designed methods. Therefore, in the future, we will explore how to fully leverage different projections and successful perspective models for 360° community.

Acknowledgement

This paper is supported by the National Natural Science Foundation of China (NSF) under Grant No. NSFC22FYT45 and the Guangzhou City, University and Enterprise Joint Fund under Grant No.SL2022A03J01278.

References

[1] Hao Ai, Zidong Cao, Jin Zhu, Haotian Bai, Yucheng Chen, and Ling Wang. Deep learning for omnidirectional vision: A survey and new perspectives. *ArXiv*, abs/2205.10468, 2022. 1

[2] 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. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13273–13282, 2023. 2, 3, 6, 8

[3] Iro Armeni, Sasha Sax, Amir Roshan Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *CoRR*, abs/1702.01105, 2017. 2, 5, 6

[4] Zidong Cao, Hao Ai, Yan Cao, Ying Shan, Xiaohu Qie, and Lin Wang. Omnizoomer: Learning to move and zoom in on sphere at high-resolution. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12851–12861, 2023. 1

[5] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *3DV*, pages 667–676. IEEE Computer Society, 2017. 1, 2, 5, 6

[6] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinghong Jiang, Feng Zhao, Bolei Zhou, and Hang Zhao. Autoalign: Pixel-instance feature aggregation for multimodal 3d object detection. In *International Joint Conference on Artificial Intelligence*, 2022. 3

[7] 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° videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1420–1429, 2018. 3

[8] Xinjing Cheng, Peng Wang, Yanqi Zhou, Chenye Guan, and Ruigang Yang. Omnidirectional depth extension networks. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 589–595, 2020. 2

[9] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1475–1483, 2017. 6

[10] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *International Conference on Machine Learning*, 2019. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *computer vision and pattern recognition*, 2009. 2, 4, 7, 8

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3, 5

[13] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12423–12431, 2019. 3

[14] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 2, 4, 5, 6, 7

[15] ChiyuMax Jiang, Jingwei Huang, Karthik Kashinath, Prabhat Prabhat, Philip Marcus, and Matthias Niessner. Spherical cnns on unstructured grids. *International Conference on Learning Representations,International Conference on Learning Representations*, 2019. 2

[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*, 6: 1519–1526, 2021. 2, 3, 4, 6, 7

[17] Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Jinhua Ma, Yaowei Wang, and Wei-Shi Zheng. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE Transactions on Multimedia*, pages 1–14, 2023. 5, 7

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 7

[19] Varun Ravi Kumar, Senthil Kumar Yogamani, Hazem Rashed, Ganesh Sitsu, Christian Witt, Isabelle Leang, Stefan Milz, and Patrick Mäder. Omnidet: Surround view cameras based multi-task visual perception network for autonomous driving. *IEEE Robotics and Automation Letters*, 6:2830–2837, 2021. 1

[20] Iro Laina, C. Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. *3DV 2016*, pages 239–248, 2016. 6

[21] Yeonkun Lee, Jaeseok Jeong, Jong Seob Yun, Wonjune Cho, and Kuk jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360° images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9173–9181, 2018. 2, 3, 4

[22] Meng Li, Senbo Wang, Weihao Yuan, Weichao Shen, Zhe Sheng, and Zilong Dong. $\mathcal{S}^2$net: Accurate panorama depth estimation on spherical surface. *IEEE Robotics and Automation Letters*, 8:1053–1060, 2023. 3

[23] Yunhao Li, Wei Shen, Zhongpai Gao, Yucheng Zhu, Guangtao Zhai, and Guodong Guo. Looking here or there? gaze following in 360-degree images. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3722–3731, 2021. 1

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 2, 3, 4, 5, 7

[25] Yunze Man, Liangyan Gui, and Yu-Xiong Wang. Bev-guided multi-modality fusion for driving perception. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21960–21969, 2023. 3

[26] Daniel Martin, Ana Serrano, Alexander W. Bergman, Gordon Wetzstein, and Belén Masiá. Scangan360: A generative model of realistic scanpaths for 360° images. *IEEE Transactions on Visualization and Computer Graphics*, 28:2003–2013, 2021. 1

[27] Rafael Monroy, Sebastian Lutz, Tejo Chalasani, and Aljoscha Smolic. Salnet360: Saliency maps for omni-directional images with cnn. *Signal Process. Image Commun.*, 69:26–34, 2017. 3

[28] Giovanni Pintore, Eva Almansa, and Jens Schneider. Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11531–11540, 2021. 1, 3, 6

[29] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360monodepth: High-resolution 360° monocular depth estimation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3752–3762, 2022. 3

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI (3)*, pages 234–241. Springer, 2015. 3, 6

[31] Mehran Shakerinava and Siamak Ravanbakhsh. Equivariant networks for pixelized spheres. *Proceedings of the 38th International Conference on Machine Learning, ICML*, abs/2106.06662, 2021. 3, 4

[32] 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*, 2022. 2, 3, 6

[33] Yu-Chuan Su and Kristen Grauman. Kernel transformer networks for compact spherical convolution. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9434–9443, 2018. 1

[34] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2573–2582, 2020. 1, 3

[35] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019. 4, 5, 7

[36] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *European Conference on Computer Vision*, 2018. 1, 2

[37] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2020. 2, 3, 4, 6

[38] 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: 5448–5460, 2022. 2, 3, 6

[39] Sirui Wang, Di Liang, Jian Song, Yuntao Li, and Wei Wu. DABERT: dual attention enhanced BERT for semantic matching. In *COLING*, pages 1645–1654. International Committee on Computational Linguistics, 2022. 6

[40] X. Wang, Ross B. Girshick, Abhinav Kumar Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2017. 4

[41] Yu yang Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2791–2800, 2022. 2, 3, 6

[42] Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guan-Sheng Wang, Xiao Tan, and Errui Ding. Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21309–21318, 2022. 1

[43] Youngho Yoon, Inchul Chung, Lin Wang, and Kuk-Jin Yoon. Spheresr: 360° image super-resolution with arbitrary projection via continuous spherical image representation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5667–5676, 2021. 1, 2, 3

[44] Haozheng Yu, Lu He, Bing Jian, Weiwei Feng, and Shanghua Liu. Panelnet: Understanding 360 indoor environment via panel representation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 878–887, 2023. 1, 3

[45] Ilwi Yun, Chan-Yong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae-Eun Rhee. Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. *ArXiv*, abs/2304.07803, 2023. 2, 3, 4, 6

[46] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. Orientation-aware semantic segmentation on icosahedron spheres. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3532–3540, 2019. 2, 4

[47] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16239–16248, 2020. 4, 7

[48] 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*, 2019. 5, 6

[49] Xu Zheng, Jinjing Zhu, Ye-Peng Liu, Zidong Cao, Chong Fu, and Lin Wang. Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1285–1295, 2023. 1

[50] Chuanqing Zhuang, Zhengda Lu, Yiqun Wang, Jun Xiao, and Ying Wang. Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. *CoRR*, abs/2112.14440, 2021. 1, 3

[51] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *ECCV (6)*, pages 453–471. Springer, 2018. 1, 2