HRDFuse: Monocular 360° Depth Estimation by Collaboratively Learning Holistic-with-Regional Depth Distributions

Hao Ai[1] Zidong Cao[1] Yan-Pei Cao[2] Ying Shan[2] Lin Wang[1,3*]
[1]AI Thrust, HKUST(GZ) [2]ARC Lab, Tencent PCG [3]Dept. of CSE, HKUST
hai033@connect.hkust-gz.edu.cn, caozidong1996@gmail.com
caoyanpei@gmail.com, yingsshan@tencent.com, linwang@ust.hk

Abstract

Depth estimation from a monocular 360° image is a burgeoning problem owing to its holistic sensing of a scene. Recently, some methods, e.g., OmniFusion, have applied the tangent projection (TP) to represent a 360° image and predicted depth values via patch-wise regressions, which are merged to get a depth map with equirectangular projection (ERP) format. However, these methods suffer from 1) non-trivial process of merging plenty of patches; 2) capturing less holistic-with-regional contextual information by directly regressing the depth value of each pixel. In this paper, we propose a novel framework, ***HRDFuse****, that subtly combines the potential of convolutional neural networks (CNNs) and transformers by collaboratively learning the holistic contextual information from the ERP and the regional structural information from the TP. Firstly, we propose a spatial feature alignment (****SFA****) module that learns feature similarities between the TP and ERP to aggregate the TP features into a complete ERP feature map in a pixel-wise manner. Secondly, we propose a collaborative depth distribution classification (****CDDC****) module that learns the* ***holistic-with-regional*** *histograms capturing the ERP and TP depth distributions. As such, the final depth values can be predicted as a linear combination of histogram bin centers. Lastly, we adaptively combine the depth predictions from ERP and TP to obtain the final depth map. Extensive experiments show that our method predicts* ***more smooth and accurate depth*** *results while achieving* ***favorably better*** *results than the SOTA methods.*

Multimedia Material

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

1. Introduction

The 360° camera is becoming increasingly popular as a 360° image provides holistic sensing of a scene with a wide field of view (FoV) [1, 4, 19, 44, 48, 52]. Therefore, the ability to infer the 3D structure of a 360° camera's surroundings has sparked the research for monocular 360° depth estimation [23, 36, 43, 45]. Generally, raw 360° images are transmitted into 2D planar representations while preserving the omnidirectional information [12, 50]. Equirectangular projection (ERP) is the most commonly used projection format [38, 49] and can provide a complete view of a scene. Cubemap projection (CP) [9] projects 360° contents into six discontinuous faces of a cube to reduce the distortion; thus, the pre-trained 2D convolutional neural networks (CNNs) can be applied. However, ERP images suffer from severe distortions in the polar regions, while CP patches are hampered by geometric discontinuity and limited FoV.

For this reason, some works [53, 54] have proposed distortion-aware convolution filters to tackle the ERP distortion problem for depth estimation. BiFuse [43] and UniFuse [23] explore the complementary information from the ERP image and CP patches to predict the depth map.

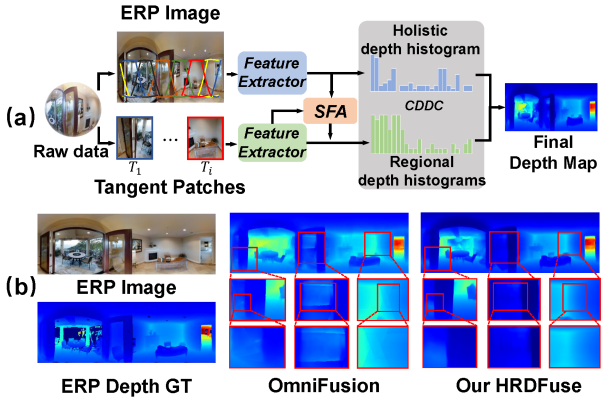

Figure 1. (a) Our HRDFuse employs the SFA module to align the regional information in discrete TP patches and holistic information in a complete ERP image. The CDDC module is proposed to estimate ERP format depth outputs from both the ERP image and TP patches based on holistic-with-regional depth histograms. (b) Compared with OmniFusion [30], our depth predictions are more smooth and more accurate.

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

Recently, research has shown that it is promising to use tangent projection (TP) because TP patches have less distortion, and many pre-trained CNN models designed for perspective images can be directly applied [16]. However, there exist unavoidable overlapping areas between two neighbouring TP patches, as can be justified by the geometric relationship in Fig. 2. Therefore, directly re-projecting the results from TP patches into the ERP format is computationally complex. Accordingly, 360MonoDepth [34] predicts the patch-wise depth maps from a set of TP patches using the state-of-the-art (SOTA) perspective depth estimators, which are aligned and merged to obtain an ERP format depth map. OmniFusion [30] proposes a framework leveraging CNNs and transformers to predict depth maps from the TP inputs and merges these patch-wise predictions to the ERP space based on geometric prior information to get the final depth output with ERP format. However, these methods suffer from two critical limitations because: 1) geometrically merging a large number of patches is computationally heavy; 2) they ignore the holistic contextual information contained only in the ERP image and directly regress the depth value of each pixel, leading to less smooth and accurate depth estimation results.

To tackle these issues, we propose a novel framework, called **HRDFuse**, that subtly combines the potential of convolutional neural networks (CNNs) and transformers by collaboratively exploring the ***holistic*** contextual information from the ERP and ***regional*** structural information from the TP (See Fig. 1(a) and Fig. 3). Compared with previous methods, our method achieves more smooth and more accurate depth estimation results while maintaining high efficiency with three key components. Firstly, for each projection, we employ a CNN-based feature extractor to extract spatially consistent feature maps and a transformer encoder to learn the depth distribution with long-range feature dependencies. In particular, to efficiently aggregate the individual TP information into an ERP space, we propose a spatial feature alignment (**SFA**) module to learn a spatially aligned index map based on feature similarities between ERP and TP. With this index map, we can efficiently measure the spatial location of each TP patch in the ERP space and achieve pixel-level fusion of TP information to obtain a smooth output in ERP format. Secondly, we propose a collaborative depth distribution classification (**CDDC**) module to learn the *holistic* depth distribution histogram from the ERP image and *regional* depth distribution histograms from the collection of TP patches. Consequently, the pixel-wise depth values can be predicted as a linear combination of histogram bin centers. Lastly, the final result is adaptive fused by two ERP format depth predictions from ERP and TP.

We conduct extensive experiments on three benchmark datasets: Stanford2D3D [2], Matterport3D [7], and 3D60 [54]. The results show that our method can achieve more smooth and more accurate depth results while favorably surpassing the existing methods by a significant margin on 3D60 and Stanford2D3D datasets (See Fig. 1 and Tab. 1). In summary, our main contributions are four-fold: (**I**) We propose HRDFuse that combines the holistic contextual information from the ERP and regional structural information from the TP. (**II**) We introduce the SFA module to efficiently aggregate the TP features into the ERP format, relieving the need for expensive re-projection operations. (**III**) We propose the CDDC module to learn the holistic-with-regional depth distributions and estimate the depth value based on the histogram bin centers.

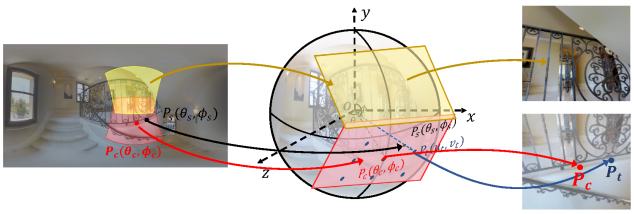

Figure 2. Geometric relationship between TP and ERP. Two TP patches are projected from the red area and yellow area.

2. Related Work

2.1. Monocular 360 Depth Estimation

ERP-based methods. To address the spherical distortion in the ERP images, endeavours have been made to leverage the characteristics of convolutional filters. OmniDepth [54] applies row-wise rectangular filters to cope with the distortions in different latitudes, while ACDNet [53] leverages a group of dilated convolution filters to rectify the receptive field. Tateno *et al*. [39] explored the standard convolution filters trained with the perspective images, and deformed the shape of sampling grids based on spherical distortion accordingly during the inference. SliceNet [33] partitions an ERP image into vertical slices and directly applies the standard convolutional layers to predict the ERP depth map.

Combination of CP and ERP. BiFuse [43] proposes to bidirectionally fuse the ERP and CP features at both encoding and decoding stages. By contrast, UniFuse [23] fuses the features only at the encoding stage as it is argued that ERP features are more important for final ERP format depth prediction. Differently, [3] employs CNNs to extract ERP features and a transformer block [14] to extract CP features, which are fused to predict the final depth map. Recently, M^3PT [46] introduces the shared random masks to process the ERP panoramas and CP depth patches simultaneously and combines the RGB information with sparse depth information to achieve panoramic depth completion.

TP-based methods. TP is recently shown to suffer less from distortion (See Fig. 2), and the pre-trained CNN models designed for perspective images can be directly applied [11]. Accordingly, 360MonoDepth [34] and OmniFusion [30] build their frameworks based on the TP patches.

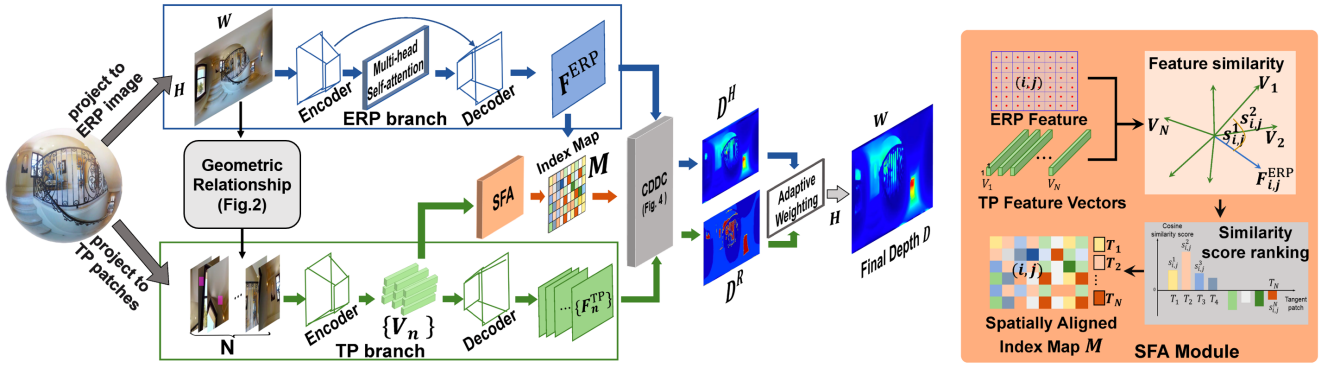

Figure 3. Overview of our HRDFuse, consisting of three parts: feature extractors for both ERP and TP inputs, spatial feature alignment (SFA) module, and collaborative depth distribution classification (CDDC) module (See Fig. 4 for details).

Concurrently, PanoFormer [36] proposes a transformer-based architecture to process the TP patches as the tokens for depth estimation. Different from [30, 34], PanoFormer designs a handcrafted sampling method to form a tangent patch by sampling eight most relevant tokens for each central token on the ERP domain rather than dividing the ERP images into TP patches. However, PanoFormer is limited by the inference speed (See Sec. 4.2), and ignores the 2D structure and spatial local information within each patch [20]. For more details, we refer readers to a recent survey [1]. In comparison, our HRDFuse combines the potential of CNNs and transformers by collaboratively learning the holistic contextual information from the ERP image and regional structural information from the TP patches.

2.2. Distribution-based Planar Depth Estimation

Many methods estimate depth by directly regressing the depth values; however, they suffer from slow convergence, and deficiency of global analysis [27, 29]. For this reason, [18] discretized the depth range into several pre-determined intervals and recast depth prediction as an ordinal regression problem, which accounts the depth distributions depending on the located intervals. Adabins [5] divides the depth range into many adaptive bins whose widths are computed from the scene information, and the depth values are a linear combination of the bin centers, showing better performance over previous methods. Our HRDFuse is the *first* to explore the idea of depth distribution classification for 360° depth estimation. The proposed CDDC module learns the holistic depth distribution histograms from the ERP image and regional depth distribution histograms from the collection of TP patches. As such, the final depth values are predicted as a linear combination of bin centers.

2.3. Vision Transformer

Transformers are capable of modeling the long-range dependencies for computer vision tasks [14, 35, 42]. Recently, it has been shown that the combination of convolutional operations and self-attention mechanisms further enhance the representation learning. For instance, DeiT [40] employs a CNN as the teacher model to distill the tokens to the transformer, while DETR [6] models the global relationship via serially feeding the features extracted by CNNs to the transformer encoder-decoder. Moreover, some works,*e.g.*, [8,32] attempted to concurrently fuse the features from CNNs and transformers. Our HRDFuse framework is also built based on the combination of CNNs and transformers; however, it shares a different spirit as we focus on ensuring network efficiency. Thus, we extract the high-resolution feature maps using a CNN-based encoder-decoder and feed them to a smaller transformer encoder [14] to estimate distributions.

3. Methodology

3.1. Overview

As depicted in Fig. 3, to exploit the complementary information from holistic context and regional structure, our framework simultaneously takes two projections of a 360° image, an ERP image and N TP patches, as inputs. For the ERP branch (See Fig. 3 Top), an ERP image with the resolution of $H \times W$ is fed into a feature extractor, comprised of an encoder-decoder block, to produce a decoded ERP feature map F^{ERP}. For the TP branch (See Fig. 3 Bottom), N TP patches are first obtained with gnomonic projection from the same sphere [16]. This indicates that the feature distributions of TP branch are closely correlated with those of the ERP branch, similar to ERP-to-CP (E2C) or C2E feature transform in [43]. Then, these TP patches are passed through the TP feature extractor to obtain 1-D patch feature vectors $\{V_n, n = 1, \ldots, N\}$, which are passed through the TP decoder to obtain the TP feature maps $\left\{F_n^{\mathrm{TP}}, n = 1, \ldots, N\right\}$.

To determine and align the spatial location of each TP patch in the ERP space and avoid complex geometric fusion for overlapping areas between neighboring TP patches, we propose the spatial feature alignment (SFA) module (Fig. 3) to learn feature correspondences between pixel vectors in the ERP feature map F^{ERP} and patch feature vectors $\{V_n\}$. This way, we can obtain the spatially aligned index map M, recording the location of each patch in the ERP space.

Next, the index map M, ERP feature map F^{ERP}, and TP feature maps $\{F_n^{\text{TP}}\}$ are fed into the proposed collaborative depth distribution classification (CDDC) module that accordingly outputs two ERP format depth predictions (See Fig. 4). In principle, the CDDC module first learns holistic-with-regional histograms to simultaneously capture depth distributions from the ERP image and a set of TP patches. Consequently, the depth distributions are then converted to depth values through a linear combination of bin centers. Lastly, the two depth predictions from the CDDC module are adaptively fused to output the final depth result. We now describe these modules in detail.

3.2. Feature Extraction

Overall, taking the ERP image and a collection of TP patches as inputs, the feature extractor of the ERP branch outputs the decoded feature map F^{ERP}, and the feature extractor of the TP branch produces encoded patch feature vectors $\{V_n\}$ and decoded TP feature maps $\{F_n^{\text{TP}}\}$.

Specifically, for the ERP branch (Fig. 3 Top), we design the feature extractor with an encoder-decoder network, following the design of OmniFusion [30]. It consists of an encoder built with the pre-trained ResNet34 [22], a multi-head self-attention block [41], and a decoder with commonly used up-sampling blocks. This way, we obtain the decoded feature map F^{ERP}.

For the TP branch, we first sample TP patches from the sphere via gnomonic projection [1, 16]. *The details can be found in the suppl. material.* Secondly, we feed the patches simultaneously into the feature extractor, similar to the ERP branch but without the multi-head self-attention block, which helps to maintain the independence of each patch feature vector for spatial feature alignment. As such, we extract the patch feature vectors $\{V_n\}$ through the encoder and obtain the decoded patch feature maps $\{F_n^{\text{TP}}\}$. The resolutions of the ERP feature map F^{ERP} and TP feature maps $\{F_n^{\text{TP}}\}$ are set to half of the corresponding input resolutions for efficiency.

3.3. Spatial Feature Alignment

With ERP feature map F^{ERP} and patch feature vectors $\{V_n\}$, our SFA module outputs the spatially aligned index map M. *It determines the spatial relations between the individual TP patches and pixel positions in the complete ERP space according to the feature similarity score ranking (See Fig. 3) and can be applied to achieve smooth pixel-wise fusion of individual TP information.* Existing works aggregate the discrete TP information into the complete ERP space via geometric fusion [30, 34]. However, they are less capable of predicting smooth equirectangular depth outputs without holistic contextual information. For instance, as shown in Fig. 1(b), depth predictions in OmniFusion suffer from severe artifacts along the edges of the merged regions. For this reason, we propose the SFA module to measure, rank, and

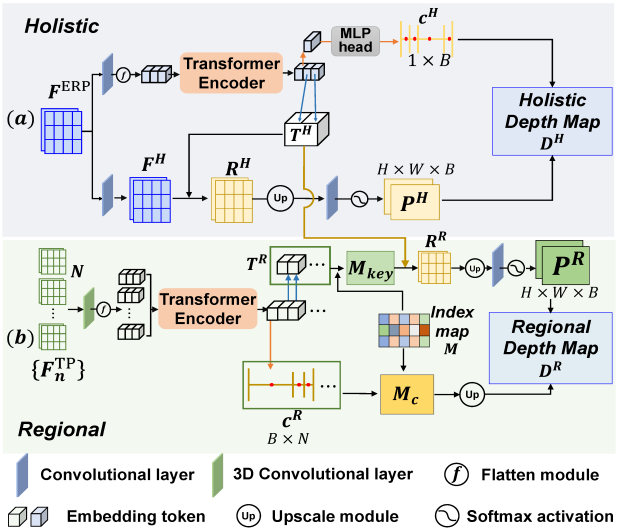

Figure 4. Overview of the CDDC module with two steps: depth distribution histogram classification, and depth prediction combination based on the range attention maps.

record the pixel-wise similarities between the ERP feature map F^{ERP} and patch feature vectors $\{V_n\}$. The pixel-wise similarity can be formulated as:

$$s_{(i,j),k} = \frac{\overrightarrow{F^{\text{ERP}}(i,j)} \cdot \overrightarrow{V_k}}{\|F^{\text{ERP}}(i,j)\| \, \|V_k\|}, \tag{1}$$

where (i, j) is the coordinate of a pixel in the ERP feature map and k is the TP patch index. As depicted in Fig. 3, for each feature vector $F^{\text{ERP}}(i, j)$ in the ERP feature map, our SFA module calculates the cosine similarity score $s_{(i,j),k}$ between $F^{\text{ERP}}(i, j)$ and each patch feature vector V_k. Then, it ranks the scores, and selects the m-th patch that satisfies:

$$m = \arg\max_k s_{(i,j),k}, \tag{2}$$

and records the index m of the pixel location (i, j) on the spatially aligned index map M. For convenience, we extend each index into an N-dimension one-hot vector and transform the resolution size of index map M to $h_e \times w_e$, where (h_e, w_e) is the resolution size of ERP feature map F^{ERP}. Note that this spatially aligned index map is produced with the guidance of the holistic contextual information only contained in the ERP image. With this index map, we can efficiently aggregate the TP features into an ERP format feature map while maintaining spatial consistency.

3.4. Collaborative Depth Distribution Classification

The proposed CDDC module replaces the pixel-wise depth value regression with depth distribution classification, inspired by the works for perspective images [5, 18]. Importantly, to fully exploit the complete view in the ERP image and structural details in the less-distorted TP patches, we

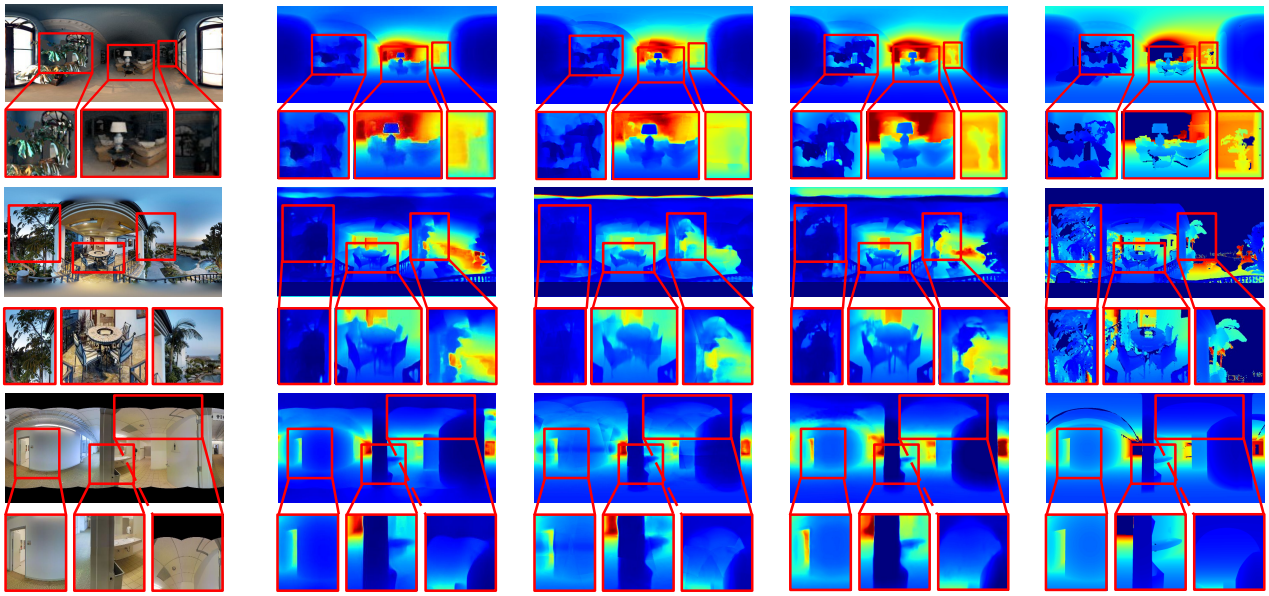

Figure 5. Qualitative results on 3D60 (top), Matterport3D (middle) and Stanford2D3D (bottom).

marry the potential of CNNs and transformers to learn the holistic-with-regional histograms capturing the ERP and TP depth distributions simultaneously. In the following, we introduce our CDDC module in three parts: generic depth distribution classification, depth prediction based on the holistic depth distribution, and depth prediction based on the regional depth distributions.

Generic depth distribution classification. Following previous works [5, 18], given an extracted feature map $F \in \mathbb{R}^{H\times W\times C_{in}}$ (*e.g.*, F^{ERP} in Fig 4(a)), a sequence of embedding tokens T_{in} is obtained from F by a convolutional layer followed by a spatially flattening module. A transformer encoder then encodes the embedding tokens T_{in}, producing processed tokens T_{out}. Note that the processed tokens T_{out} now benefit from the global context and thus can accurately capture the depth distribution. Then the first token $T_{out}[0]$ from T_{out} is selected to predict the bin centers $\mathbf{c}$ of depth distribution histograms (*e.g.*, $\mathbf{c}^H$ in Fig 4(a)) as:

$$\mathbf{c}_i = D_{min} + (\mathbf{w}_i/2 + \sum_{j=1}^{i-1} \mathbf{w}_j), \tag{3}$$

$$\mathbf{w}_i = (D_{max} - D_{min}) \frac{(\text{mlp}(T_{out}[0]))_i + \epsilon}{\sum_{j=1}^{B} (\text{mlp}(T_{out}[0]))_j + \epsilon}, \tag{4}$$

where $i, j = 1, \ldots, B$, $\mathbf{w}$ is the bin widths of the distribution histogram, mlp denotes a multi-layer perceptron (MLP) head with a ReLU activation, (D_{min}, D_{max}) is the depth range of the dataset, B denotes the number of depth distribution bins, and ϵ is a small constant to ensure that each value of $\mathbf{w}$ is positive. Finally, the bin centers $\mathbf{c}$ are linearly blended with a probability score map P (*e.g.*, P^H in the Fig 4(a)) to predict the depth value at each pixel (i, j):

$$D(i, j) = \textstyle\sum_{b=1}^{B} P(i, j)_b \cdot \mathbf{c}_b. \tag{5}$$

Holistic distribution-based depth prediction. As depicted in Fig. 4(a), we follow the process of generic depth distribution classification to predict the holistic depth bin centers $\mathbf{c}^H$. We then perform the following steps to obtain the holistic probability score map P^H. First, we select a part of processed tokens, which are the output of the transformer encoder and contain global context, as the "query" embedding T^H. At the same time, we encode a spatially consistent feature map F^H containing local pixel-wise information as the "keymap". Next, we calculate the dot-production between the query T^H and pixel features in F^H to obtain a range attention map R^H. This range attention map R^H thus contains global context and is spatially aligned with the ERP feature map. Then R^H is passed through a 1×1 convolutional layer with a softmax activation to predict the probability score map P^H. Given holistic depth bin centers and probability score map, we can now calculate the holistic depth map following Eq. 5. Note that the ERP feature map is with the half resolution of the input ERP image to limit GPU memory usage. Therefore, we additionally employ an up-sampling module to upscale the probability score map to the desired resolution (i.e., $H \times W$).

Regional distribution-based depth prediction. Compared with the ERP branch, predicting an ERP format depth map from TP patches based on corresponding regional depth distributions meets two critical difficulties: 1) accurate and smooth fusion of individual TP patches; 2) capturing the holistic information for the ERP format depth output. To address them, we utilize the spatially aligned index map M

Datasets	Method	Patch size/FoV	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE(log) ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Stanford2D3D	FCRN [28]	−/−	0.1837	-	0.5774	-	0.7230	0.9207	0.9731
	BiFuse with fusion [43]	−/−	0.1209	-	0.4142	-	0.8660	0.9580	0.9860
	UniFuse with fusion [23]	−/−	0.1114	-	0.3691	-	0.8711	0.9664	0.9882
	OmniFusion (2-iter) [30]	256 × 256 / 80°	0.0950	0.0491	0.3474	0.1599	0.8988	0.9769	0.9924
	PanoFormer* [36]	−/−	0.1131	0.0723	0.3557	0.2454	0.8808	0.9623	0.9855
	HRDFuse,Ours	128 × 128 / 80°	0.0984	0.0530	0.3452	0.1465	0.8941	0.9778	0.9923
	HRDFuse,Ours	256 × 256 / 80°	**0.0935**	0.0508	**0.3106**	**0.1422**	**0.9140**	**0.9798**	**0.9927**
Matterport3D	FCRN [28]	−/−	0.2409	-	0.6704	-	0.7703	0.9714	0.9617
	BiFuse with fusion [43]	−/−	0.2048	-	0.6259	-	0.8452	0.9319	0.9632
	UniFuse with fusion [23]	−/−	0.1063	-	0.4941	-	0.8897	0.9623	0.9831
	OmniFusion (2-iter) * [30]	256 × 256 / 80°	0.1007	0.0969	0.4435	0.1664	0.9143	0.9666	0.9844
	PanoFormer* [36]	−/−	0.0904	0.0764	0.4470	0.1650	0.8816	0.9661	0.9878
	HRDFuse,Ours	128 × 128 / 80°	0.0967	0.0936	**0.4433**	**0.1642**	**0.9162**	**0.9669**	0.9844
	HRDFuse,Ours	256 × 256 / 80°	0.0981	0.0945	0.4466	0.1656	0.9147	0.9666	0.9842
3D60	FCRN [28]	−/−	0.0699	0.2833	-	-	0.9532	0.9905	0.9966
	Mapped Convolution [15]	−/−	0.0965	0.0371	0.2966	0.1413	0.9068	0.9854	0.9967
	BiFuse with fusion [43]	−/−	0.0615	-	0.2440	-	0.9699	0.9927	0.9969
	UniFuse with fusion [23]	−/−	0.0466	-	0.1968	-	0.9835	0.9965	0.9987
	ODE-CNN [10]	−/−	0.0467	0.0124	0.1728	0.0793	0.9814	0.9967	0.9989
	OmniFusion (2-iter) [30]	128 × 128 / 80°	0.0430	0.0114	0.1808	0.0735	0.9859	0.9969	0.9989
	HRDFuse,Ours	128 × 128 / 80°	0.0363	0.0103	0.1565	0.0594	0.9888	**0.9974**	0.9990
	HRDFuse,Ours	256 × 256 / 80°	**0.0358**	**0.0100**	**0.1555**	**0.0592**	**0.9894**	0.9973	**0.9990**

Table 1. Quantitative comparison with the SOTA methods. $*$ represents that the model is re-trained following the official setting. **red** indicates that our method achieves the best performance.

from the SFA module and the holistic query embedding T^H from the ERP branch (See Fig. 4(b)). We first follow the generic depth distribution classification to collect regional depth bin centers from the collection of TP feature maps $\{F_n^{\mathrm{TP}}\}$ and concatenate them to obtain the tensor $\mathbf{c}^R$ with the size $B \times N$. Then, with the spatial guidance of index map M, we can obtain an ERP format bin center map M_c from bin center vector set $\mathbf{c}^R$ as:

$$M_c(i,j) = \sum_{n=1}^{N} M(i,j)_n \cdot \mathbf{c}_n^R \tag{6}$$

where (i, j) is the pixel coordinate, and n is the patch index. The bin center map M_c represents the depth distribution of each pixel with aggregated regional structural information. Meanwhile, we concatenate and average a collection of processed regional tokens, which record the regional structural information of each individual TP patch, to a tensor T^R. Similarly, the index map M then helps to aggregate the regional structure in T^R to a regional feature map M_{key}. Next, with M_{key} as the "keymap" and T^H as the "query", we can predict the regional probability score map P^R and further output the ERP format regional depth map D^R. Note that the query embedding T^H from the ERP branch provides necessary and favorable holistic guidance. *Due to the page limit, more details,* e.g.*, network architecture, can be found in Table. 1 of the suppl. material.*

3.5. The Final Output and Loss Function

To obtain the final depth map, we adaptively fuse the depth prediction D^H from the holistic contextual branch and depth prediction D^R from the regional structural branch, which can be formulated as follows:

$$D = w_0 D^H + w_1 D^R, \tag{7}$$

where w_0 and w_1 are learnable parameters and $w_0 + w_1 = 1$ (superiority of adaptive weighting is shown in Table. 6). Following previous works [23, 30], we adopt BerHu loss [27] for pixel-wise depth supervision, denoted as $\mathcal{L}_{depth}$. Furthermore, to encourage the holistic distribution to be consistent with all depth values in the ground truth depth map, we adopt the commonly used bi-directional Chamfer loss [17] as the holistic distribution loss $\mathcal{L}_{H_{bin}}$. Therefore, the total loss $\mathcal{L}_{total}$ can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_{depth} + \lambda \mathcal{L}_{H_{bin}}, \tag{8}$$

where λ is a weight factor and set to 0.1 for all experiments empirically [5].

4. Experiments

Datasets and Metrics. We conduct experiments on three benchmark datasets: Stanford2D3D [2], Matterport3D [7], and 3D60 [54]. Note that Stanford2D3D and Matterport3D are real-world datasets, while 3D60 is composed of two synthetic datasets (SUNCG [37] and SceneNet [21]) and two real-world datasets (Stanford2D3D and Matterport3D). However, there exists a issue in the 3D60 dataset, which is mentioned by UniFuse [23] that the problematic rendering may cause some problems with depth prediction.

Following previous works [23, 30, 43], we evaluate our method with the standard metrics: Absolute Relative Er-

ERP branch	TP branch	geometric fusion	SFA	CDDC	FPS	#Params	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
✓					**7.88**	**33.57M**	0.1028	0.0985	0.4543	0.9086	0.9658	0.9841
	✓	✓			3.56	37.09M	0.1018	0.0982	0.4492	0.9104	0.9662	0.9842
✓	✓	✓			2.82	70.66M	0.0986	0.0944	0.4466	0.9141	0.9664	0.9843
✓	✓		✓		6.21	49.95M	0.0991	0.0956	0.4479	0.9132	0.9666	0.9843
✓	✓	✓		✓	3.23	56.96M	0.0978	0.0940	0.4458	0.9146	0.9666	0.9841
✓	✓		✓	✓	5.52	53.77M	**0.0967**	**0.0936**	**0.4433**	**0.9162**	**0.9669**	**0.9844**

Table 2. The ablation results for individual components. Both ERP and TP branch are trained with the depth distributions following [5].

Number	Patch size/FoV	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑
10	128×128 / 80°	0.0996	0.0965	0.4491	0.9130	0.9664
18		**0.0967**	**0.0936**	**0.4433**	**0.9162**	0.9669
26		0.0978	0.0945	0.4444	0.9151	**0.9670**
46		0.1232	0.1178	0.4996	0.8780	0.9563
10	256×256 / 80°	0.0976	0.0948	0.4447	0.9152	0.9668
18		0.0981	0.0945	0.4466	0.9147	0.9666
26		0.0974	0.0953	0.4478	0.9147	0.9662
46		**0.0966**	**0.0938**	**0.4432**	**0.9168**	**0.9668**

Table 3. The ablation results for the number of TP patches.

Patch FoV	Patch size	Abs Rel ↓	Sq Rel ↓	RMSE ↓
60	128×128	**0.0986**	0.0961	0.4454
	256×256	0.0986	**0.0942**	**0.4448**
80	128×128	**0.0967**	**0.0936**	**0.4433**
	256×256	0.0981	0.0945	0.4466
100	128×128	**0.0970**	**0.0938**	**0.4453**
	256×256	0.0979	0.0940	0.4458

Table 4. The ablation results for the TP patch size and FoV.

ror (Abs Rel), Squared Relative Error (Sq Rel), Root Mean Squared Error (RMSE), Root Mean Squared Logarithmic Error (RMSE (log)), as well as a percentage metric with a threshold δ_t, where $t \in \{1.25^1, 1.25^2, 1.25^3\}$. *Due to the lack of space, the details of datasets and metrics can be found in the suppl. material.*

Implementation Details. We implement our method using Pytorch and train it on a single NVIDIA 3090 GPU. We use ResNet-34 [22], pre-trained on ImageNet [13], as the encoder. Following [30], we use Adam [24] optimizer with cosine annealing [31] learning rate policy and set the initial learning rate to 10^{-4}. The default TP patch number is $N = 18$, and the batch size is 4. Following [30], we train 80 epochs for Stanford2D3D [2] and 60 epochs for Matterport3D [7], and 3D60 [54]. The input images are augmented only by horizontal translation and horizontal flipping.

4.1. Comparison with the state-of-the-arts

In Table. 1, we compare our HRDFuse with the SOTA methods on three benchmark datasets. For a fair comparison, we do not discuss self-supervised methods [25, 26, 41]. Note that OmniFusion did not provide the pre-trained models on the Matterport3D dataset, thus we re-trained them with the official hyper-parameters. PanoFormer did not provide the experiment details, *e.g.*, epochs; thus for fair comparison, we re-trained the model with the same setting using the official hyper-parameters for the same epochs. For all the datasets, we show the results of the proposed HRDFuse with TP patch sizes of 128×128 and 256×256.

As shown in Table. 1, our HRDFuse **performs favorably against** the SOTA methods [23, 28, 30, 43, 54] **by a significant margin on two of the three datasets.** Specifically, for the Stanford2D3D dataset, our HRDFuse with the patch size of 256×256 outperforms UniFuse [23] by 16.07% (Abs Rel), 15.85% (RMSE), and 4.29% (δ_1). Compared with OmniFusion (2-iter), our HRDFuse improves RMSE(log) by 11.07% and δ_1 by 1.52%. *More comparisons with it can be found in the suppl. material due to the space limit.*

For Matterport3D and 3D60 datasets, which contain more samples, our HRDFuse is more advantageous and surpasses the compared methods for all metrics. On the Matterport3D dataset, our HRDFuse with the patch size 128×128 outperforms UniFuse by 2.65% (δ_1), 9.03% (Abs Rel), outperforms PanoFormer by 3.46% (δ_1) and outperforms OmniFusion (2-iter) by 3.97%(Abs Rel), 3.41% (Sq Rel). On the 3D60 dataset, HRDFuse with the patch size 256×256 outperforms UniFuse by 23.18% (Abs Rel) and 20.99% (RMSE), and outperforms OmniFusion (2-iter) by 16.74% (Abs Rel) and 13.99% (RMSE).

In Fig. 5, we present the qualitative comparison with UniFuse 5(b) and OmniFusion 5(c). Our HRDFuse can recover more regional structural details (e.g., leaves and seats) and suffer less from artifacts caused by the discontinuity among TP patches (red boxes). *More qualitative comparisons can be found in the suppl. material.*

4.2. Ablation Study and Analyses

The effectiveness of each module. We verify the effectiveness of each module in our HRDFuse by adding one module each time (Table. 2). We form our baselines in three ways. Firstly, for the ERP branch-only baseline, we directly follow the Adabins [5] to predict the holistic depth distributions from the ERP images and regress the depth maps. Secondly, with only the TP branch, we add the geometric fusion, as done in [30], to the feature extractor to obtain the ERP format depth map. Thirdly, we combine the ERP branch and TP branch, followed by the geometric fusion mechanism in [30]. Based on this, we then add the SFA module. Here, we directly leverage the spatially aligned index map to aggregate the patch feature vectors V_n into an ERP feature map and predict the depth map, without em-

Number of bins	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑
20	0.0997	0.0963	0.4502	0.9132	0.9661
50	0.0971	0.0939	0.4454	0.9159	0.9665
100	**0.0967**	**0.0936**	**0.4433**	**0.9162**	**0.9669**
150	0.0997	0.0948	0.4497	0.9121	0.9662

Table 5. Impact of the number B of depth histogram bins.

ERP branch	TP branch	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑
1	0	0.0976	0.0948	0.4450	0.9153	0.9664
0	1	0.0975	0.0944	0.4459	0.9149	0.9670
0.5	0.5	0.0969	0.0942	0.4442	0.9157	0.9668
Adaptive weighting		**0.0967**	**0.0936**	**0.4433**	**0.9162**	**0.9669**

Table 6. The ablation study for the final fusion.

ploying the decoder (see Fig. 3) or geometric fusion module of [30] in the TP branch. Lastly, we add the CDDC module to learn the holistic-with-regional depth distributions.

As shown in Table. 2, with the ERP branch alone, it is difficult to alleviate the projection distortion, thus leading to the worst depth estimation performance. The performance improves when using the TP branch only due to less distortion, and is further improved by the fusion of the ERP branch and TP branch (with the geometric fusion mechanism). Furthermore, by introducing the SFA module, the network parameters are significantly reduced by 29.31%, leading to more than three frames per second (FPS) gain in inference speed. When the CDDC module is finally added, the performance is further boosted by 2.42%(Abs Rel) and 2.09%(Sq Rel), although the parameters slightly increase. Especially, compared with PanoFormer (20.37 M parameters), our method higher FPS (5.52) than it (4.93).

Patch size, FoV, and the number of patches of TP. They are essential parameters and directly affect the accuracy and efficiency of our method. Thus, we study their impact and find an optimal balance between efficiency and performance. Following [30], we fix the patch number as 18 and examine how TP patch size affects the learning under multiple patch FoVs. As in Table 4, on the Matterport3D dataset, all the results with the patch size of 128×128 perform better than those of 256×256, which indicates that too large patch size may cause the redundancy of regional structural information and degrade the accuracy of the final ERP output. Meanwhile, we can observe the influence of patch FoV in Table 4: either too small patch FoV or too large patch FoV degrades the performance. When FoV is too small, the regional information in each TP patch would be insufficient; in contrast, too large FoV increases the inconsistency in the overlapping areas between adjacent TP patches.

Furthermore, as the number of TP patches and the computational memory cost are directly related, we fix the patch size and FoV to compare the depth results with different patch numbers such that we can find the most cost-effective patch number. As shown in Table. 3, too few patches can not provide sufficient region-wise structural details, while too many patches lead to the redundancy of details, thus degrading the role of holistic contextual information. We find that $N = 18$ performs best in our experiments.

Number of bins. We now compare the performance with various numbers of depth distribution histogram bins. As observed from Table. 5, starting from $B = 20$, the depth accuracy first improves with the increase of B, and then drops significantly. The result indicates that too many bins lead to difficulty in classification. For this reason, we choose 100 as the number of bins for experiments.

Weights of fusion. Table. 6 lists the depth results under 4 groups of fusion weights with the patch number set as $N = 18$, patch size as 128×128, and FoV as $80°$. Overall, our adaptive weighting achieves the best performance.

Rationality of SFA module. As depicted in Fig. 2 and Table. 2, the geometric fusion of [30] requires more inference time to ensure the depth values of overlapping areas among TP patches. By contrast, our SFA module can provide the alignment in the feature space, which is more efficient and effective. As shown in *Fig. 7 in the suppl. material*, when the holistic scene structure is simple, SFA module makes the index map (*Fig. 7(b)*) centralized to several representative TP patches with higher frequency of index (*e.g.*, index 4, 6 in *Fig. 7(c)*) to avoid redundant usage. This indeed validates the overlap among the TP patches, as shown in Fig. 2. In comparison, when the scene structure becomes more complex (*Fig. 8 in the suppl. material*), more TP patches (with index 12,16 in *Fig. 8(c)*) are needed to describe the holistic depth information.

5. Conclusion and Future Work

This paper proposed a novel solution for monocular 360° depth estimation, which predicts an ERP format depth map by collaboratively learning the holistic-with-regional depth distributions. To address the two issues: 1) challenges in pixel-wise depth value regression; 2) boundary discontinuities brought by the geometric fusion, our HRDFuse introduced the SFA module and the CDDC module, whose contributions allow HRDFuse to efficiently incorporate ERP and TP, and significantly improve the depth prediction accuracy and obtain favorably better results. Our work focused on the supervised monocular 360° depth estimation and did not cover self-supervised methods. In the future, we will explore the potential of TP, e.g., contrastive learning for TP patches. In addition, our task and 360° semantic segmentation [47,51] are closely related, as they are both dense scene understanding tasks. Therefore, joint 360° monocular depth estimation and semantic segmentation based on the combination of ERP and TP is a promising research direction.

Acknowledgement

This work was supported by the CCF-Tencent Open Fund and the National Natural Science Foundation of China (NSFC) under Grant No. NSFC22FYT45.

References

[1] Hao Ai, Zidong Cao, Jinjing Zhu, Haotian Bai, Yucheng Chen, and Lin Wang. Deep learning for omnidirectional vision: A survey and new perspectives. *CoRR*, abs/2205.10468, 2022. 1, 3, 4

[2] 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, 6, 7

[3] Jiayang Bai, Shuichang Lai, Haoyu Qin, Jie Guo, and Yanwen Guo. Glpanodepth: Global-to-local panoramic depth estimation. *CoRR*, abs/2202.02796, 2022. 2

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5460–5469, 2022. 1

[5] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, pages 4009–4018. Computer Vision Foundation / IEEE, 2021. 3, 4, 5, 6, 7

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV (1)*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. 3

[7] 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. 2, 6, 7

[8] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. *CoRR*, abs/2108.05895, 2021. 3

[9] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. Cube padding for weakly-supervised saliency prediction in 360° videos. In *CVPR*, pages 1420–1429. Computer Vision Foundation / IEEE Computer Society, 2018. 1

[10] Xinjing Cheng, Peng Wang, Yanqi Zhou, Chenye Guan, and Ruigang Yang. Omnidirectional depth extension networks. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 589–595, 2020. 6

[11] Benjamin Coors, Alexandru Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *ECCV*, 2018. 2

[12] H. S. M. Coxeter. Introduction to geometry. 1961. 1

[13] 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. 7

[14] 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 *ICLR*. OpenReview.net, 2021. 2, 3

[15] Marc Eder, True Price, Thanh Vu, Akash Bapat, and Jan-Michael Frahm. Mapped convolutions. *ArXiv*, abs/1906.11096, 2019. 6

[16] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *CVPR*, pages 12423–12431. Computer Vision Foundation / IEEE, 2020. 2, 3, 4

[17] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. *computer vision and pattern recognition*, 2016. 6

[18] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011. Computer Vision Foundation / IEEE Computer Society, 2018. 3, 4, 5

[19] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7174–7182, 2019. 1

[20] Jianyuan Guo, Kai Han, Han Wu, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. Cmt: Convolutional neural networks meet vision transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12165–12175, 2022. 3

[21] Ankur Handa, Viorica Patraucean, Simon Stent, and Roberto Cipolla. Scenenet: An annotated model generator for indoor scene understanding. *international conference on robotics and automation*, 2016. 6

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 4, 7

[23] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics Autom. Lett.*, 6(2):1519–1526, 2021. 1, 2, 6, 7

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[25] Weifeng Kong, Qiudan Zhang, You Yang, Tiesong Zhao, Wenhui Wu, and Xu Wang. Self-supervised indoor 360-degree depth estimation via structural regularization. In *PRICAI*, 2022. 7

[26] Ziye Lai, Dan Chen, and Kaixiong Su. Olanet: Self-supervised 360° depth estimation with effective distortion-aware view synthesis and l1 smooth regularization. *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 7

[27] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. *international conference on 3d vision*, 2016. 3, 6

[28] Iro Laina, C. Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. *2016 Fourth International Conference on 3D Vision (3DV)*, pages 239–248, 2016. 6, 7

[29] Sihaeng Lee, Janghyeon Lee, Byungju Kim, Eojindl Yi, and Junmo Kim. Patch-wise attention network for monocular depth estimation. In *AAAI*, pages 1873–1881. AAAI Press, 2021. 3

[30] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. *CoRR*, abs/2203.00838, 2022. 1, 2, 3, 4, 6, 7, 8

[31] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv: Learning*, 2016. 7

[32] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *ICCV*, pages 357–366. IEEE, 2021. 3

[33] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. Slicenet: Deep dense depth estimation from a single indoor panorama using a slice-based representation. In *CVPR*, pages 11536–11545. Computer Vision Foundation / IEEE, 2021. 2

[34] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360monodepth: High-resolution 360° monocular depth estimation. *CoRR*, abs/2111.15669, 2021. 2, 3, 4

[35] Wenqi Shao, Yixiao Ge, Zhaoyang Zhang, Xuyuan Xu, Xiaogang Wang, Ying Shan, and Ping Luo. Dynamic token normalization improves vision transformer. *arXiv preprint arXiv:2112.02624*, 2021. 3

[36] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360° depth estimation. In *ECCV*, 2022. 1, 3, 6

[37] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *arXiv: Computer Vision and Pattern Recognition*, 2016. 6

[38] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360° imagery. In *NIPS*, 2017. 1

[39] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *ECCV (16)*, volume 11220 of *Lecture Notes in Computer Science*, pages 732–750. Springer, 2018. 2

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 2021. 3

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 7

[42] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Ge Yuying, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 3

[43] 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 *CVPR*, pages 459–468. Computer Vision Foundation / IEEE, 2020. 1, 2, 3, 6, 7

[44] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Led2-net: Monocular 360° layout estimation via differentiable depth rendering. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12951–12960, 2021. 1

[45] 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*, PP, 2022. 1

[46] Zhiqiang Yan, Xiang Li, Kun Wang, Zhenyu Zhang, Jun Yu Li, and Jian Yang. Multi-modal masked pre-training for monocular panoramic depth completion. *ECCV*, 2022. 2

[47] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1376–1386, 2021. 8

[48] Sheng Yang, Beichen Li, Yan-Pei Cao, Hongbo Fu, Yu-Kun Lai, Leif Kobbelt, and Shi-Min Hu. Noise-resilient reconstruction of panoramas and 3d scenes using robot-mounted unsynchronized commodity rgb-d cameras. *ACM Transactions on Graphics (TOG)*, 39(5):1–15, 2020. 1

[49] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3358–3367, 2019. 1

[50] Youngho Yoon, Inchul Chung, Lin Wang, and Kuk-Jin Yoon. Spheresr: 360deg image super-resolution with arbitrary projection via continuous spherical image representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5677–5686, 2022. 1

[51] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16896–16906, 2022. 8

[52] Xinyu Zhang, Yao Zhao, Nikk Mitchell, and Wensong Li. A new 360 camera design for multi format vr experiences. *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 1273–1274, 2019. 1

[53] 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, 2

[54] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *ECCV (6)*, volume 11210 of *Lecture Notes in Computer Science*, pages 453–471. Springer, 2018. 1, 2, 6, 7