Useg-PanoDepth:Unified 360° Depth Estimation for Indoor and Outdoor Scenes with Semantic Assistance

Qingling Chang, Jingheng Xu, Yan Cui, Yikui Zhai, Senior Member, IEEE, Pasquale Coscia, Senior Member, IEEE, Angelo Genoves, Senior Member, IEEE, Vincenzo Piuri, Fellow, IEEE, and Fabio Scotti, Senior Member, IEEE

Abstract—In complex 360° scenes, depth estimation is challenging for small objects and the depth of object boundaries, which cannot be effectively solved with existing works. 360° depth estimation is unable to produce uniform depth estimate findings in both indoor and outdoor settings due to the datasets. In this paper, the Useg-PanoDepth and PanoDepth dataset is proposed to improve the above problems effectively. The Diagonal-aware Attention Module (DAM) effectively estimates small objects in complex scenes. Enhanced Boundary Module (EBM), for enhancing boundary information,can also effectively solve the problem of depth unification of indoor and outdoor scenes. Extensive experiments on our constructed PanoDepth dataset, Useg-PanoDepth achieves SOTA results. The Relative accuracy ($delta < 1.25$) reaches 87.82%. Following is the link to the source code at https://github.com/xjh6/Useg-PanoDepth.

Index Terms—360° Depth Estimate, 360° Depth Dataset, Indoor and outdoor universal, Segmentation embeding, Small object and boundary enhenced.

I. Introduction

DEPTH estimate plays a crucial part in 3D reconstruction. The 360° image contains the entire surrounding environment (360° × 180°) information, which plays an important role in three-dimensional reconstruction. Therefore, after depth estimation for the 360° image, three-dimensional reconstruction can be performed to reconstruct the entire scene. Currently, the depth estimation of ordinary images has been deeply studied, and methods based on deep learning have obtained excellent results [1], [2]. However, there are relatively few studies on depth estimation of 360° images with equirectangular projection, and the research results have not yet reached the level of depth estimation of ordinary images [3].

After equirectangular projection (ERP) processing, 360° images have certain image distortion problems. There have been many studies on image distortion problems in the past to reduce the accuracy impact caused by image distortion. Regarding image distortion, researchers mainly deal with it from two directions: 1. by improving the feature extraction approach and altering the convolution in CNN [4]–[6], seeking to increase the accuracy of the CNN model in this way. [7] [8] blends the benefits of Transformer and CNN to create TEDepth, which has demonstrated exceptional performance; 2. by using additional cubemap or other data-assisted methods, some researchers [9], [10] use cubemap projection (CMP), another projection method of 360° image, to assist the 360° monocular depth estimation model to combat image distortion. [11] using semantic information to assist in guiding depth estimation edges to improve the accuracy of depth estimation for object boundaries. There are also some researchers [12] using self-supervised learning to improve model performance that is affected by the insufficient number of 360° image datasets.

The shortcomings of current research on depth estimation of 360° images are: 1. For the estimation accuracy of small objects and object boundary information in 360° images, the depth of some small objects is often misestimated. This paper proposes a DAM module using the attention mechanism to calculate the attention of local positions, trying to focus on local information to alleviate the situation where the depth of small objects is easily misestimated in 360° image depth estimation. This paper also uses the semantic segmentation results output by SAM [13] as the mask and proposes the EBM module to fuse the rough depth map with the mask to further retain small objects in complex scenes and improve

This study was funded by Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation. ("Climbing Program" Special Funds.)(No. pdjh2023b0530), Construction of Knowledge Graph for Science and Technology Resource Big Data(No. 2019AL032), AI Neural Network Algorithm Endows Optical Cameras with Laser Capability(No. 2021ZDJS095), Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011576), Guangdong Higher Education Innovation and Strengthening School Project (No. 2020ZDZX3031, No. 2022ZDZX1032, No. 2023ZDZX1029), Wuyi University Hong Kong and Macao Joint Research and Development Fund (No. 2022WGALH19), Guangdong Jiangmen Science and Technology Research Project (No. 2220002000246, No. 2023760300070008390). (Corresponding author: Yikui Zhai)

Qingling Chang,Jingheng Xu,and Yikui Zhai is with the School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, China (e-mail: qlchangcas@gmail.com; e-mail: juxheg@gmail.com;e-mail: yikuizhai@163.com).

Yan Cui is with the School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, China and Zhuhai 4Dage Network Technology, Zhuhai 519000, China (e-mail: cuiyan@wyu.edu.cn)

Pasquale Coscia, Angelo Genovese, Vincenzo Piuri and Fabio Scotti are with the Department of Computer Science and Dipartimento di Informatica, Universita Degli Studi di Milano, 20133, Milan, Italy (e-mail: pasquale.coscia@unimi.it; angelo.genovese@unimi.it; vincenzo.piuri@unimi.it; fabio.scotti@unimi.it).

the accuracy of object edge depth estimation. 2. 360° image depth estimation cannot obtain uniform results for indoor and outdoor scene estimation. The reason is also affected by the 360° image dataset [14], [15]. Real-world datasets are essential because training the model on synthetic datasets would result in poor depth estimate generalization in the real world, even if the issue of a shortage of data for 360° images depth estimation has been mitigated by the emergence of synthetic datasets. Currently, there is a lack of datasets with 360° images of both indoor and outdoor scenes. To this end, this paper constructs the PanoDepth dataset. It has a 360° image dataset of both indoor and outdoor scenes. Numerous 360° images in complicated situations, such as utility rooms and fire scenes, are included in this PanoDepth collection. This makes depth estimation for 360° images even more challenging. The information that follows is an overview of our contributions:

1. We constructed the PanoDepth dataset, which consists of 3,189 pairs of RGB-D images, including both complex indoor and outdoor scenes. The dataset contains 1,135 RGB-D image pairs for outdoor scenarios and 2,054 pairs for indoor environments.

2. We propose the Diagonal-aware Attention Module (DAM) to alleviate the problem of losing object depth in 360° image depth estimation in complex scenes through local attention calculation.

3. We propose the Enhanced Boundary Module (EBM) to better utilize the mask obtained by the semantic segmentation branch to fuse with the rough depth map. Simultaneously, the mask error is fixed, and the 360° image's object boundary estimation is further refined.

4. We trained Useg-PanoDepth. It surpasses the performance of existing work in PanoDepth. Also in Pano3D the Relative accuracy ($\delta < 1.25$) reaches 94.56%.

The remainder of this paper is organized as follows: Section II reviews the related literature; Section III provides a detailed description of the PanoDepth dataset; Section IV introduces the proposed Useg-PanoDepth algorithm; Section V presents and analyzes the experimental results, comparing them with state-of-the-art (SOTA) 360° depth estimation methods; Section VI discusses the potential application scenarios of the proposed approach; and Section VII concludes the paper with a summary of the findings.

II. Related work

A. 360° Monocular Depth Estimation

In 3D reconstruction, depth estimation is an essential step that 360° image aid in restoring in areas with weak textures. There are two main methods for processing 360° images: 1. Equirectangular Projection (ERP), 2. Cubemap Projection (CMP). ERP allows a single image to cover the entire scene, while CMP requires six images from different perspectives to cover the entire scene, as shown in the Fig. 1.

Make3D [16] is the first work on depth estimation and 3D reconstruction on a single image. However, it is depth estimation using traditional methods. Numerous deep learning-based depth estimation methods have been developed as a result of this [5], [17]–[21] of deep learning. Among them, they approach the challenge of depth estimation as either a problem involving regression or a problem with classification. OmniDepth [20] is the first work to propose using RectNet for depth estimation of ERP 360° images, which performs better than using separately processed stereoscopic mapping (CMP) from different angles. However, OmniDepth cannot handle the distortion of geometric structures well, so many existing methods are dedicated to solving this problem.

(a) Equirectangular Projection(ERP) (b) Cubemap Projection(CMP)

Fig. 1. Compare with the Equirectangular Projection and the Cubemap Projection. The two types of projections are as follows: (a) Equirectangular Projection (ERP), which can cover the full scene with one image but will distort it; and (b) Cubemap Projection (CMP), which covers the complete scene with many images at different positions.

First, some work deals with the distortion of ERP by changing the convolution kernel to expand the receptive field. For example, SphereNet [22] calculates the sampling position of the inverse noise. The convolution between the longitude and latitude angles and the spherical field of view is defined by CFL [23]. Furthermore, some techniques add new CMP branches. BiFuse [9], for example, develops a bi-projection fusion module to achieve feature fusion and extracts features for ERP and CMP using two identical encoder-decoder structures, respectively. In addition, UniFuse [10] proposes a more efficient fusion method that removes the decoder in the CMP branch. Methods using the CMP branch suffer from the inability to extract good features of weakly textured areas. Some recent works have been dedicated to using unsupervised, self-supervised [24], [25] methods to overcome the lack of 360° images data. And they are utilizing joint learning strategies to combine self-supervised and fully supervised methods. BGDNet [36] first infers background depth from walls, floors and ceilings using background masks, room layout, and camera models.

Some researchers believe that the CNN network cannot solve the problem of image distortion well. Nowadays, networks with transformers as backbone have swept through all major basic computer vision tasks. Its excellent performance has brought new directions to many researchers. Ranftl et al. [26] applied vision transformer to dense prediction tasks for the first time, and proposed three different transformer architectures to achieve excellent performance in dense prediction tasks of computer vision. Based on 360° customization, Shen et al. [27] created the initial depth estimation monocular transformer. It

TABLE I
Comparison of different 360° image dataset.

Dataset Name	Pano3D [15]	Structured3D [14]	Stanfor2d3d [38]	PanoDepth
Images Number	7907	196k	1413	3189
Data Type	Real	Synthetic	Real	Real
Scenes Type	In	In	In	In & Out
RGB Resolution	1024×512	1024×512	1024×512	1024×512
Depth Resolution	1024×512	1024×512	1024×512	1024×512
Depth Range(m)	$(0, 8)$	N/A	$(1/512, 128)$	$(1/256, 100)$

Fig. 2. Four types of images in our dataset. (a) Input RGB image. (b) Raw depth map acquired by laser camera. (c) The mask information generated by SAM and the subsequent mask image obtained through manual correction. (d) Use depth completion technology to convert the original depth map into a dense depth map that can be used as ground truth.

Fig. 3. The first column shows some images from the Pano3D dataset, which are from real-world images. The red boxes mark the parts of the image with poor quality. The second column shows some images from the Structured3D dataset, which are synthetic images. The third row displays a portion of the Stanford2D3D [38] dataset. However, this subset contains a limited number of 360° images. The fourth column shows some images from our collected dataset, which are from real-world scenes.

successfully lessens the effects of images distortion.

B. 360° Datasets

Scene modeling with 360° images has progressively gained popularity in recent years. Additionally, the development of deep learning-based techniques has been aided by the release of many 360° datasets. The Pano3D [15] datasets is currently the most widely used indoor real-world 360° datasets. And the Structured3D [14] is the synthetic dataset. The 360° KITTI [28] dataset is an actual outdoor road dataset that has a wide range of annotations. It may be used for a variety of applications, including object detection, semantic segmentation, depth estimation, and more. Most of the current research methods are solely focused on depth estimation of indoor or outdoor scenes. Therefore, this paper proposes a new 360° image dataset that integrates indoor and outdoor complex scenes in the real world. We named it PanoDepth dataset.

III. PanoDepth Dataset

All 360° images in the PanoDepth dataset are from the real world. It has numerous indoor and outdoor 360° scenes. Especially in some complex scenes containing 360° images, the depth estimation method of images brings greater challenges.

Our dataset consists of RGB images and depth maps captured using the LiDAR and the 4DMega[1] camera. The LiDAR has a wavelength of 905 nm, a point frequency of 200,000 points per second, and an accuracy of ±0.02m. The 4DMega camera lens has a horizontal field of view (FOV) of 133.11° and a vertical FOV of 85.06°. The resolution of the images captured by the camera is 5472×3648 pixels, while the resolution of the panoramic images, synthesized after rotational capture, is $16,384 \times 8,192$ pixels, as shown in Fig. 2 (a). The dataset includes 11 types of scenes, such as exhibition halls, factories, office environments, and residential areas. The depth data range is $(0, 100m)$, and the depth error is ±0.01m. In order to reduce the storage space occupied by the dataset and reduce a certain amount of calculation, we follow the practice of the conventional 360° dataset and reset the 360° image and depth map to 1024×512 size, saving a lot of storage space, greatly reducing the amount of calculation during model training, and making the trained model achieve higher accuracy.

According to our research on the Panodepth, although the original depth map data is denser than the depth information obtained by other devices, as shown in Fig. 2 (c), it still cannot meet the conditions as ground truth in deep learning. Therefore, we employ the depth completion technology [29] to finish the depth of the initial depth map data, and obtain more dense depth data as shown in Fig. 2 (d). By observing our original depth image data, we found that there are many small holes (missing data) in the data. The distance from the outdoor scene to

[1]https://www.4dkankan.com/#/Mega

the sky is infinite, and we set its depth to zero, which is consistent with the missing data at the upper and lower edges of the depth map data. Therefore, we use the depth completion method ip-basic [30] that does not require training and uses computer graphics technology to fill in missing depth values. The [30] applies five morphological operations to depth maps. First, depth values are inverted using $D_{inverted} = 120.0 - D_{input}$ to preserve object boundaries during dilation. Dilation fills empty pixels, followed by a closing operation to fill gaps and connect object boundaries. Small to medium holes are filled through dilation, while large holes are inferred from neighboring depths. Denoising is done with a 5×5 median filter, and smoothing is applied with a 5×5 Gaussian filter. Finally, the original depth encoding is restored using $D_{out} = 120.0 - D_{inverted}$. This process enhances the depth map's completeness and accuracy. The final result is shown in Fig. 2 (d).

To quantify the inherent errors in depth completion, we performed a comparative error analysis between the original sparse depth map and the completed depth map. Specifically, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were computed by aligning and evaluating non-zero pixels in the sparse depth map against their corresponding pixels in the completed depth map. The resulting error metrics yielded an MAE of 0.2329m and an RMSE of 0.9094m, which fall within acceptable tolerance thresholds, thereby validating the practical reliability of the depth completion framework.

We divided the 3189 images into 7:2:1 training, test, and validation sets. And keep the ratio of indoor and outdoor data in each subset the same as the PanoDepth dataset. There are 1135 RGB-D image pairings for outside situations and 2054 pairs for inside scenes. We broadly categorize the scenes in the PanoDepth dataset into 11 types, including exhibition halls, office environments, fire scenes, factories, and more. There are 7 categories of outdoor scenes, such as factory perimeters, parks, and residential areas. Additionally, there are 8 categories of indoor scenes, including factory interiors, office spaces, exhibition halls, and fire scenes.

The Pano3D dataset is an improvement over the 3D60 dataset, resolving the issue of depth information leakage caused by light sources in the latter. The Structured3D dataset features a large number of indoor 360° images, with data sourced from the virtual world, and provides highly accurate depth information. The Stanford2D3D dataset includes a smaller set of 360° images, with depth data covering both the top and bottom of the images.

According to Table I, the PanoDepth dataset in this paper is smaller than the Structured3D and Pano3D datasets, but larger than the Stanford2d3d [38] dataset. All image scene types in the Pano3D, Stanford2d3d, and Structured3D datasets are from indoor scenes. In contrast, the scene types in our PanoDepth dataset cover both indoor and outdoor environments. The PanoDepth dataset matches the other three in terms of 360° image resolution. Regarding depth data, the PanoDepth dataset falls between the two.

Another crucial point is that the Structured3D dataset is created using synthetic data. As a result, it contains a vast collection of images and precise depth maps. However, it is challenging to transfer models trained on synthetic datasets to the real world. The data in this study, on the other hand, is derived from real-world sources. Although it is smaller, it is still crucial for 360° data.

It can be observed from Fig. 3 that the image quality of PanoDepth is better than the other two datasets. Especially in the area marked by the red box, the 360° images in the Pano3D dataset are distorted and missing. Since the data in the Structured3D collection is synthetic, there aren't any visible mistakes or omissions. Stanford2D3D also contains high-quality image data, but it has a limited number of 360° images, with only 1413 available.

IV. Method

A. Overview

The entire architecture is displayed in Fig. 4. It consists of a backbone, convolution decoder, DAM, SAM[10], EBM, and output header module. The branch obtains the 360° segmentation mask of the RGB image through SAM, and the main branch performs supervised learning by inputting the RGB image and depth ground truth map. Unify all input images into 1024×512 sizes and input them to the backbone to obtain four types of feature maps, t he size of which is the original size 1/32,1/16,1/8,1/4, and process it through convolution decoder. Following these four feature maps, the four feature maps are fused together using the DAM and the EBM to create an approximate depth map that is 256×128 in size. And then fused with the mask obtained by the branch through the EBM, and finally through The output head gets the final fine-tuned predicted depth map.

B. Diagonal-aware Attention Module

In the context of panoramic images and 3D reconstruction, diagonal pixels often carry distinctive depth variations or geometric features that are essential for accurate depth estimation. In traditional attention mechanisms, the model typically attends to all pixels uniformly, without explicitly considering spatial relationships that might exist in specific regions of the image. We hypothesized that focusing attention on diagonal pixels could help the model better capture important geometric structures, which are often significant in the 360° images tasks.

The network can autonomously determine which parts of the image need attention because of the attention mechanism. As the Fig. 4 (b) shown.We offer a Diagonal-aware Attention Module to enhance the accuracy of object detail in depth estimate by better balancing global and local information. In the Non-Local Network [31], it can calculate the interaction between any two locations to directly capture long-range dependencies without being limited to adjacent points. CCNet [32] is developed from [31]. The latter calculates attention by focusing on

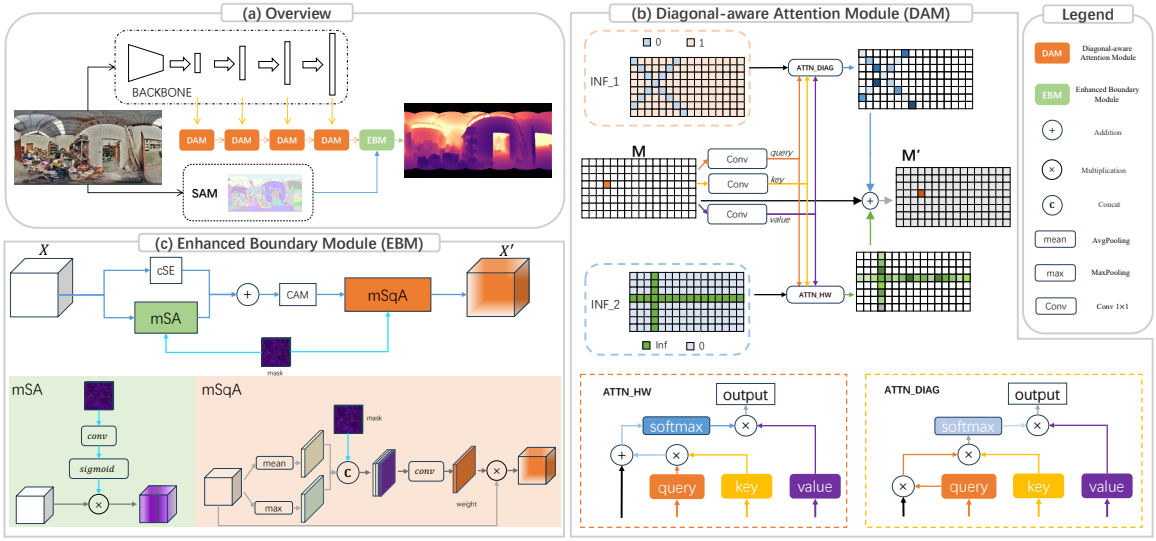

Fig. 4. (a) is an overview of the entire model. The DPT model, which combines CNN with Transformer feature extraction, is used by Backbone. (b) The four feature maps that the backbone extracted are combined by the Diagonal Aware Attention Module (DAM). In order to optimize the precision of local depth value prediction, the technique entails masking the adjacent pixels of every pixel on the main, secondary, and vertical diagonals, in that order. After calculating the attention in each of these directions, the initial value and the attention score are combined. (c) To create a fine depth map, the Enhanced Boundary Module (EBM) combines the fused feature map produced by the EBM with the segmentation map produced by the SAM branch. In order to include the segmentation output from the SAM branch into the coarse depth feature map, this module employs two stages of spatial attention.

horizontal and vertical pixels to reduce the calculation amount of the former. Because the amount of information covered by 360° images is much larger than that of ordinary images. Therefore, estimating the depth of 360° images in complex scenes requires more attention to local details. In an effort to raise the level of focus in the surrounding areas even further. Based on CCNet, we added the calculation of pixels on the diagonal. This allows us to associate all adjacent pixels in a certain pixel with one calculation. Experiments have demonstrated that our approach produces excellent outcomes.

Fig. 4 (b) illustrates the diagonal-aware attention module in detail. Among them, M is the input feature map. $M \in \mathbb{R}^{C\times H\times W}$ undergoes three one-by-one convolutions to obtain $query \in \mathbb{R}^{3H\times H\times W}$, $key \in \mathbb{R}^{3H\times H\times W}$, $value \in \mathbb{R}^{C\times H\times W}$. The query, key and value are used for attention calculation.

For a 360° image with height H, the length of the diagonals varies depending on the location of the pixels. Therefore, we classify the diagonals based on their lengths. A 360° image with height H corresponds to H+1 different diagonal lengths. We use INF_1 to create a tensor composed of values 1 and 0, and denote it as $A \in \mathbb{R}^{C^{'}\times H\times W}$. The value of $C^{'}$ is three times of H.

The *query*, *key* and *value* obtained after convolving A and the input M are calculated in ATTN_DIAG. We perform the following operations on query and A:

$$Q_{i,j} = query_{i,j} \cdot A_{i,j} \tag{1}$$

where i=[1,2,...,H],j=[1,2,...,W]. $A_{i,j}$ is the specific position [i,j] of the pixel in a certain layer channel in A. The same is true for $query_{i,j}$ and $Q_{i,j}$. The $A_{i,j}$, $query_{i,j}$, $Q_{i,j} \in \mathbb{R}^{C^{'}}$. This is a pixel-by-pixel multiplication of query and A of the same size. The $Q_{i,j}$ can mask the value in the diagonal direction of the pixel and is the key to calculating diagonal attention. For Q, *key* and *value*, tensor reconstruction becomes $Q^{'} \in \mathbb{R}^{HW\times C^{'}}$, $key^{'} \in \mathbb{R}^{C^{'}\times HW}$, $value^{'} \in \mathbb{R}^{C\times HW}$. Then $Q^{'}$, $key^{'}$, $value'$ calculate the attention as follows:

$$ATTN_{DIAG} = value^{'} \cdot softmax(Q^{'} \cdot key^{'}) \tag{2}$$

where $ATTN_{DIAG} \in \mathbb{R}^{C\times HW}$ is the attention score in the diagonal direction of the pixel. And by reconstructing the tensor dimensions, the same size $DIAG \in \mathbb{R}^{C\times H\times W}$ as M is obtained.

Similarly, for pixels in the horizontal and vertical directions, we refer to the method used in CCNet [32]. We use INF_2 to create a tensor composed of values Inf and 0, and denote the result of the computation between INF_2 and M as HW. In order to automatically modify the attentional influence, we further employ a learnable parameter weight ϕ. When connected with the residual structure, the final DAM is defined below:

$$M^{'} = \phi \cdot (DIAG + HW) + M \tag{3}$$

where $M^{'} \in \mathbb{R}^{C\times H\times W}$. $M^{'}$ has the same dimensions as the input M, but $M^{'}$ associates each pixel with its diagonal and horizontal and vertical directions. The experiments have demonstrated that DAM may significantly raise the focus on specific regions in the images and enhance the accuracy of depth estimation.

C. Enhanced Boundary Module

We will describe how to obtain segmentation masks in the next subsection. However, we found that segmentation

masks have certain shortcomings. First, there are some wrong segmentation masks in the segmentation results. Secondly, the obtained segmentation results will distinguish patterns on the surface of certain objects, such as wall decorations, etc. Direct fusion of the initially obtained segmentation mask and the predicted depth map is likely to lead to incorrect depth estimation. As a way to increase the estimation accuracy even more, we therefore suggested that the Enhanced Boundary Module combine the feature map and mask using a two-stage fusion method.

The $mask \in \mathbb{R}^{1\times H\times W}$ we get from the SAM branch, we first perform convolution and other operations on it to perform preliminary processing of different categories in the mask, and then broadcast the input. The weights are passed to each layer channel of the feature map X. We name this operation mSA, and its definition is as follows:

$$X_1 = X \cdot sigmoid(conv(mask)) \tag{4}$$

where $X, X_1 \in R^{C\times H\times W}$. The series of operations on the mask are to change the values of different categories in the mask and generate a weight to increase attention to the object boundary information. In addition, we also calculate the channel attention cSE [33] for X, and use the Squeeze operation and the Excitation operation to correct the feature channel information. Then add it to mSA to get an $X_2 \in \mathbb{R}^{C\times H\times W}$, which is the first stage of fusion operation and is defined as follows:

$$X_2 = cSE + X_1 \tag{5}$$

Subsequently, we use the Channel Attention Module [34] operation on the first-stage fusion result A channel-corrected feature map X_3 is obtained, and then the second stage of fusion is performed. We average and maximize X_3 across channels to obtain X_{mean}, $X_{max} \in \mathbb{R}^{1\times H\times W}$ respectively, and stack X_{mean}, X_{max} , $mask$ on the channel. And the fusion operation is implemented through MLP to obtain an $X_{weight} \in \mathbb{R}^{1\times H\times W}$, which is defined as follows:

$$X_{weight} = MLP(concat(X_{mean}, X_{max}, mask)) \tag{6}$$

To further enhance the model's focus on edge information, we implemented the second step of fusion operations using the broadcast technique once more. The final definition is as follows:

$$X^{'} = X_{weight} \cdot X_3 \tag{7}$$

where $X^{'} \in \mathbb{R}^{C\times H\times W}$. Experiments have proven that EBM ignores the error information that may be contained in the mask output by the SAM branch and corrects the rough depth map. Additionally, to fully utilize the SAM transfer capability's zero-shot capacity in order to lower the loss of small objects and increase the accuracy of object edges in 360° depth estimation.

V. Experiments

A. Implementation Details

To achieve a balance between accuracy, computational efficiency, and memory footprint, we adopt an input resolution of 1024×512 for model training. To mitigate potential alignment errors in ground truth data and ensure evaluation integrity, our analysis is restricted to the vertically cropped central region of the 360° image, spanning pixel heights 128 to 384. The architecture employs DPT [26] as the backbone network. Training is conducted using the Adam optimizer with momentum parameters $\beta = (0.5, 0.999)$, a fixed learning rate of 2×10^{-4}, and a batch size of 1. The model undergoes 70 training epochs on our dataset without employing data augmentation techniques.

B. Evaluation Quantitative

We assess our model's performance using the six measures that have been proposed in earlier work. Here is a definition of the six error metrics:

Absolute Relative Difference (AbsRel):

$$AbsRel = \frac{1}{|N|}\sum y \in N|g_{pred} - g_t|/g_t \tag{8}$$

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{|N|}\sum_{y\in N}(||g_{gpred} - g_t||)^2} \tag{9}$$

Mean absolute error (MAE):

$$MAE = \frac{1}{|N|}\sum y \in N|g_{pred} - g_t| \tag{10}$$

Threshold(δ):

$$\delta_k = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left(\max\left(\frac{d_i^{\text{pred}}}{d_i^{\text{gt}}}, \frac{d_i^{\text{gt}}}{d_i^{\text{pred}}}\right) < 1.25^k\right) \tag{11}$$

where $k = 1, 2, 3$, d_i^{pred} denotes the predicted depth value for the i-th pixel, and d_i^{gt} represents the corresponding ground-truth (GT) depth value.

C. Quantitative Comparison

We conduct a quantitative comparative analysis of different 360° depth estimation models on two datasets. Table II shows the data results. First, in our PanoDepth dataset, the depth values range from 0 to 100 meters. Therefore, our evaluation of the models in [35], [10], [27], [12], and [37] requires adjusting the range of depth values predicted by these models. Second, we only evaluate the middle part of the 360° image, especially the area between the image heights [128, 384]. This is because the ground truth (GT) data in our dataset lacks depth information in the top and bottom regions of the image. In previous studies, researchers usually adopt this approach by removing the top and bottom regions of the image for evaluation. We also train and evaluate all models in the same experimental environment.

The quantitative comparison in the PanoDepth is displayed in Table II. Our method surpasses existing 360° depth estimation methods in six indicators. In terms of error metrics, our method demonstrates significant

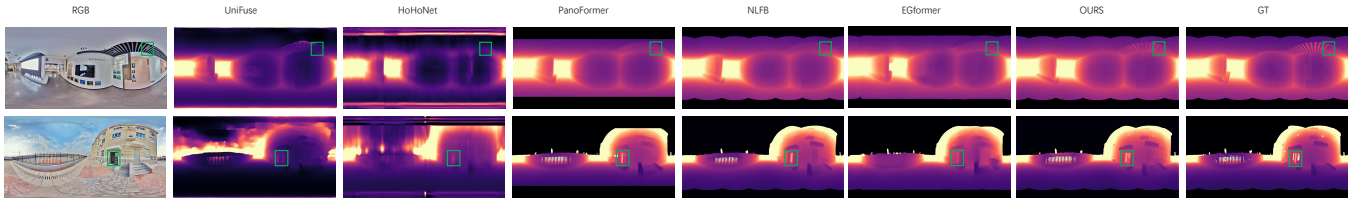

Fig. 5. The qualitative comparison with different models. The green boxes indicate the variations in how different models estimate the same little object. The depth distance of small objects can be accurately estimated by our model. For instance, our model provides an accurate estimate for the lights in the first column and the gaps in the ceiling.

TABLE II
Quantitative comparison in the PanoDepth

Method	Error metric↓			Accuracy metric↑		
	Abs Rel	RMSE	MAE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
HoHoNet [35]	0.1721	2.1537	0.8182	0.8078	0.9262	0.9655
UniFuse [10]	0.1610	2.2520	0.7573	0.8176	0.9295	0.9669
Panoformer [27]	0.1444	2.4744	0.8349	0.8500	0.9393	0.9689
NLFB [12]	0.1289	1.6710	0.6274	0.8727	0.9492	0.9743
EGformer [37]	0.2321	2.5428	1.1030	0.7355	0.8930	0.9475
Ours	0.1253	1.6305	0.6027	0.8782	0.9513	0.9754

TABLE III
Quantitative comparison in the Pano3D

Method	Error metric↓			Accuracy metric↑		
	Abs Rel	RMSE	MAE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
HoHoNet [35]	0.1164	0.4782	0.2411	0.9063	0.9703	0.9844
UniFuse [10]	0.1098	0.3512	0.2144	0.8804	0.9733	0.9900
NLFB [12]	0.0871	0.5972	0.2982	0.9170	0.9751	0.9897
Panoformer [27]	0.0664	0.3091	0.1422	0.9436	0.9847	0.9939
EGformer [37]	0.0785	0.4113	0.2200	0.9363	0.9817	0.9926
EGformer*[37]	0.0660	0.3281	0.1699	0.9502	0.9877	0.9952
Ours	0.0579	0.3119	0.1437	0.9456	0.9841	0.9933

* The asterisk (*) indicates results evaluated under the EGformer [37], which calculates metrics over the entire 360° image without applying the vertical crop ([128, 384] pixels) used by baseline methods.

improvements over previous methods. In the Abs Rel index, our method achieves a value of only 0.1253, which is 2.79% better than the state-of-the-art (SOTA) result, while HoHoNet is even higher at 0.1721. The closest competitor, NLFB [12], has a value of 0.1289. In RMSE, our method scores 1.6305, which is 2.42% better, while Panoformer reaches 2.4744. The closest NLFB to our method is 1.6710. For MAE, our method scores 0.6027, which is 3.93% better, while Panoformer reaches 0.8349, and the closest NLFB is 0.6274. As for the recent work EGformer, it did not perform well on the PanoDepth dataset.

In terms of accuracy, our method has also made significant progress. Under the ($\delta < 1.25$) metric, our method achieves 0.8782, while HoHoNet only reaches 0.8078. The second-best method, NLFB, has a score of 0.8727. Our method also performs well in other cases. Although EGformer is a recent work, its accuracy performance is unsatisfactory.

Most research on 360° depth estimation is conducted in indoor scenes, which necessitates transferring the technology to outdoor scenes. This transfer may be affected by scenarios where the depth value is 0, such as in outdoor scenes with large areas of sky. Methods like [10], [12], [27], [37] and [35] struggle to handle outdoor distant scenes effectively, resulting in lower evaluation results.

We evaluate our approach on the publicly available indoor 360° dataset Pano3D to confirm its superiority even further. The assessment information is displayed in Table III.

Our method also achieves quite competitive performance on the Pano3D dataset. In the image height range of [128,384], the accuracy index ($\delta < 1.25$) reaches 94.56%, which is better than all models. This further shows that our method is effective and effectively improves the estimation accuracy of small objects in 360° depth estimation by using DAM and EBM.

D. Qualitative Comparison

To examine in more detail the impact of our approach on depth estimation in comparison to other approaches like [10] [12] [27] [35] and [37]. We visualize the predicted depth of the image, as shown in Fig. 5. We use green boxes to mark some details of the 360° image so that we can easily see the advantages of our method and the methods of [10] [12] [27] [35] and [37]. Provide additional evidence of our method's efficacy in estimating the depth of small objects.

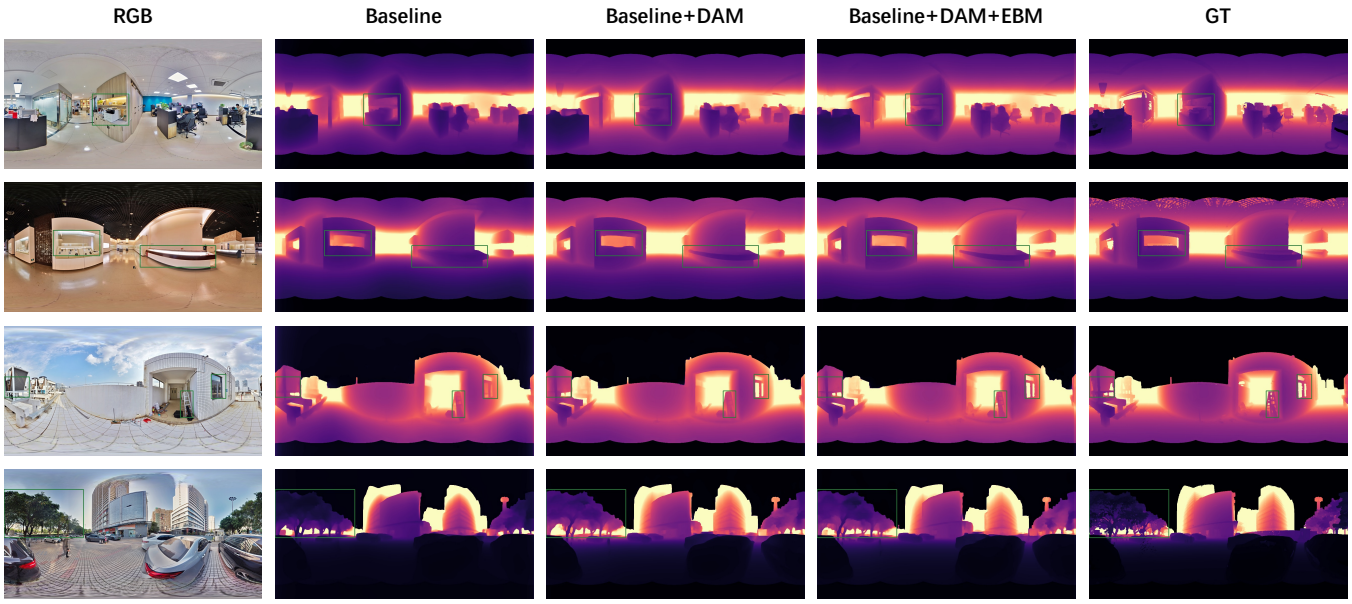

Fig. 6. The qualitative comparison with ablation study. According to the results of the ablation experiment, DAM has a slightly better depth estimation of small objects than the Baseline. After adding EBM, the depth of small objects in the gaps between objects will be better. The green box in the figure can be used for horizontal comparison. The module we proposed can effectively improve the accuracy of depth estimation.

TABLE IV
Ablation study in the PanoDepth

baseline	DAM	EBM	Error metric↓			Accuracy metric↑		
			Abs Rel	RMSE	MAE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
√			0.1326	1.6767	0.6311	0.8718	0.9488	0.9739
√	√		0.1303	1.6781	0.6228	0.8743	0.9493	0.9739
√		√	0.1302	1.6788	0.6266	0.8739	0.9491	0.9739
√	√	√	0.1253	1.6305	0.6027	0.8782	0.9513	0.9754

We present two distinct situations, one inside and one outside, in Fig. 5. While the details of NLFB [12] are very rough and do not display the shape of the ceiling cracks and walls, our approach can clearly estimate the different depths as in GT in the details of the exhibition hall ceiling cracks and light bulbs in the inside scene. As an illustration, consider the three-dimensional lettering and ceiling on the left side of the image and the wall border on the right. The spaces between the wall and the ceiling are still not as obvious as ours, despite UniFuse and us being compared. Compared to existing approaches, our method can establish crisper boundaries in outdoor situations for things like jutting air conditioners.

E. Ablation Study

We performed ablation studies to more thoroughly assess the impact of the Enhanced Boundary Module (EBM) and the Diagonal-aware Attention Module (DAM). Table IV displays the outcomes of the experiment. When added to the baseline, the DAM we provided can achieve a level similar to the NLFB model. What is special is that NLFB also uses additional dataset for training.

As shown in Fig. 6, we conduct a qualitative analysis of the ablation study. We also use green boxes to mark details. We analyze 4 scenarios. There are two indoor scenes and two outdoor scenes. Among them, for protruding objects and recessed showcases in indoor scenes, DAM can be more accurate than Baseline estimation. Combined with EBM, the object boundary accuracy can be improved and some scenarios where both baseline and DAM can estimate errors can be corrected. Two outdoor scenes, such as trees. The baseline cannot effectively estimate the depth level of leaves, and the performance of DAM is slightly better than Baseline. After the improvement of EBM, the different depth levels of leaves can be more clearly distinguished.

In Conclusion, We further improved the estimation accuracy of our model after adding EBM. As a result, the DAM and EBM that we suggested can both significantly raise the model's estimation accuracy.

TABLE V
Compare with FPS

with local SAM	with SAM	FPS
√		0.378
	√	0.099

F. Efficiency Analysis of Adding SAM by Inference

Due to the operational efficiency, the split mask used for training in this paper is output in advance using SAM and saving the split mask results locally for reading. This method's sole goal is to increase the model's training efficiency. Therefore, this subsection will discuss the efficiency comparison between using reading the pre-saved-to-local segmentation mask results during inference and simulates actual use which without a pre-generated segmentation mask during inference.

We employ the read pre-stored local split mask as a control group to provide a fair comparison. Instead, we remove the SAM module. As shown in Table V, when simulating normal usage, the inference FPS of our model is at 0.099. When using the pre-stored segmentation mask, the inference FPS of our model is at 0.378. Although the time to use the SAM branch will be longer when simulating normal usage.

VI. Application

Depth estimation is a very critical task in 3D reconstruction. Results from high-precision depth estimation can raise the 3D reconstruction's accuracy. Efficiency and accuracy in 3D reconstruction tasks have always been its shortcomings. Our dataset includes some information gathered from the fire scene. Through high-precision depth estimation and three-dimensional reconstruction, the fire scene can be well reconstructed. It can help firefighters quickly and safely find the cause of the fire. With the advancement of virtual reality technology, online house viewing has gradually become a trend. High-precision three-dimensional reconstruction allows consumers to view houses immersively anytime and anywhere. Similarly, there are a large number of museum scenes in the dataset. Three-dimensional reconstruction is performed through accurate depth estimation to create online tours, helping people who cannot get there to see and learn. As shown in Fig. 7.

Unfortunately, due to the difficulty in obtaining 360° data, the number of our dataset is not enough and the proportion of indoor and outdoor scenes is unbalanced, so we cannot perform better in outdoor scenes.

VII. Conclusion

This paper proposes a branch-assisted 360° depth estimation model based on the mask provided by SAM to improve accuracy and achieve uniform indoor and outdoor accuracy. On our PanoDepth dataset, we outperform all open-source 360° image depth estimation models. We also proposed DAM and EBM, both of which are plug-and-play attention modules. Our dataset has been the subject of numerous tests to demonstrate the efficacy of this module.

Eventually, we're going to work on expanding the dataset to perform depth completion in a more accurate and effective way to provide more accurate depth data as ground truth (GT) increase the precision of the depth estimation model. Expanding more 360° images of indoor and outdoor scenes can also improve the estimation accuracy. Based on the images, we discovered that, like the sky, the 3D reconstruction of outside sceneries is not very good and is influenced by the depth at infinity. We will further improve the depth estimation accuracy of outdoor scenes and increase the number of outdoor scenes in the dataset. Secondly, we will do further research on the distortion characteristics of 360° images and strive to overcome the influence of object deformation brought by 360° images to increase the model's accuracy even more. Finally, due to limitations in the acquisition equipment, the collected depth images lack coverage of the top and bottom regions. We plan to address this limitation by improving the completeness of the depth map ground truth (GT), ensuring that the upper and lower surfaces are included, thereby enhancing the overall integrity of the panoramic depth. We will also conduct further research on the different depth distribution effects brought by indoor and outdoor scene images, reduce the error caused by the infinite depth of outdoor images, and use a segmentation model with higher accuracy and more efficiency to assist in enhancing depth estimation precision.

At present, the accuracy of this method has achieved the expected best effect, but its reasoning efficiency still needs to be improved. We will also focus on lightweight models and improving reasoning efficiency, so that the model can be used in more application scenarios and bring more convenient usage features.

References

[1] Kondapaneni, N., Marks, M., Knott, M., Guimaraes, R., Perona, P.: Text-image alignment for diffusion-based perception. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13883–13893 (2024),

[2] Yang, Lihe and Kang, Bingyi and Huang, Zilong and Xu, Xiaogang and Feng, Jiashi and Zhao, Hengshuang.: Depth anything: Unleashing the power of large-scale unlabeled data. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.pp. 10371–10381 (2024)

[3] da,Silveira, T.L., Pinto, P.G., Murrugarra-Llerena, J., Jung, C.R.: 3d scene geometry estimation from 360 imagery: A survey. ACM Computing Surveys 55(4), 1–39 (2022)

[4] Tateno, K., Navab, N., Tombari, F.: Distortion-aware convolutional filters for dense prediction in panoramic images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 707–722 (2018)

[5] Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems 27 (2014)

[6] Su, Y.C., Grauman, K.: Kernel transformer networks for compact spherical convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9442–9451 (2019)

[7] Shao, Shuwei, et al. "Towards comprehensive monocular depth estimation: Multiple heads are better than one." IEEE Transactions on Multimedia 25 (2022): 7660-7671.

[8] Shao, S., Pei, Z., Chen, W., Li, R., Liu, Z., Li, Z. (2023). Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimation. IEEE Transactions on Multimedia.

[9] Wang, F.E., Yeh, Y.H., Sun, M., Chiu, W.C., Tsai, Y.H.: Bifuse: Monocular 360 depth estimation via bi-projection fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 462–471 (2020)

[10] Jiang, H., Sheng, Z., Zhu, S., Dong, Z., Huang, R.: Unifuse: Unidirectional fusion for 360 panorama depth estimation. IEEE Robotics and Automation Letters 6(2), 1519–1526 (2021)

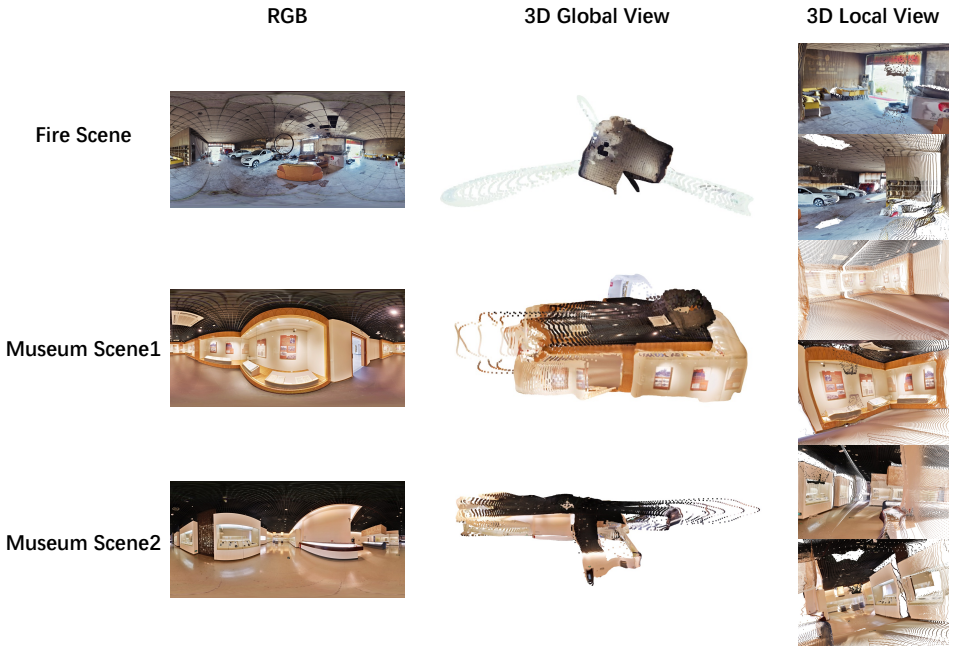

Fig. 7. 3D reconstruction is performed via the depth map output by our model. Among them, overview is the overall structure of the three-dimensional reconstruction, and local is the internal scene details. Fire Scene is the structure restored from the 360° image after the fire, Museum1 and Museum2 are the structures of the exhibition hall.

[11] Zhu, S., Brazil, G., Liu, X.: The edge of depth: Explicit constraints between segmentation and depth. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13116–13125 (2020)
[12] Yun, I., Lee, H.J., Rhee, C.E.: Improving 360 monocular depth estimation via non-local dense prediction transformer and joint supervised and self-supervised learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3224–3233 (2022)
[13] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
[14] Zheng, Jia, et al. "Structured3d: A large photo-realistic dataset for structured 3d modeling." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer International Publishing, 2020.
[15] Albanis, G., Zioulis, N., Drakoulis, P., Gkitsas, V., Sterzentsenko, V., Alvarez, F., Zarpalas, D., Daras, P.: Pano3d: A holistic benchmark and a solid baseline for 360 depth estimation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3722–3732. IEEE (2021)
[16] Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence 31(5), 824–840 (2008)
[17] Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
[18] Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016)
[19] Wofk, D., Ma, F., Yang, T.J., Karaman, S., Sze, V.: Fastdepth: Fast monocular depth estimation on embedded systems. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 6101–6108. IEEE (2019)
[20] Zioulis, N., Karakottas, A., Zarpalas, D., Daras, P.: Omnidepth: Dense depth estimation for indoors spherical panoramas. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 448–465 (2018)
[21] Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4009–4018 (2021)
[22] Coors, B., Condurache, A.P., Geiger, A.: Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In: Proceedings of the European conference on computer vision (ECCV). pp. 518–533 (2018)
[23] Fernandez-Labrador, C., Facil, J.M., Perez-Yus, A., Demonceaux, C., Civera, J., Guerrero, J.J.: Corners for layout: End-to-end layout recovery from 360 images. IEEE Robotics and Automation Letters 5(2), 1255–1262 (2020)
[24] Zioulis, N., Karakottas, A., Zarpalas, D., Alvarez, F., Daras, P.: Spherical view synthesis for self-supervised 360 depth estimation. In: 2019 International Conference on 3D Vision (3DV). pp. 690–699. IEEE (2019)
[25] Zhou, K., Wang, K., Yang, K.: Padenet: An efficient and robust panoramic monocular depth estimation network for outdoor scenes. In: 2020 ieee 23rd international conference on intelligent transportation systems (itsc). pp. 1–6. IEEE (2020)
[26] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
[27] Shen, Z., Lin, C., Liao, K., Nie, L., Zheng, Z., Zhao, Y.: Panoformer: Panorama transformer for indoor 360° depth estimation. In: European Conference on Computer Vision. pp. 195–211. Springer Nature Switzerland Cham (2022)
[28] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
[29] Hu, J., Bao, C., Ozay, M., Fan, C., Gao, Q., Liu, H., Lam, T.L.: Deep depth completion from extremely sparse data: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(7), 8244–8264 (2022)
[30] Ku, J., Harakeh, A., Waslander, S.L.: In defense of classical image processing: Fast depth completion on the cpu. In: 2018 15th Conference on Computer and Robot Vision (CRV). pp. 16–22. IEEE (2018)
[31] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)

[32] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 603–612 (2019)
[33] Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. pp. 421–429. Springer (2018)
[34] Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
[35] Sun, C., Sun, M., Chen, H.T.: Hohonet: 360 indoor holistic understanding with latent horizontal features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2573–2582 (2021)
[36] Chen, J., Wan, Z., Narayana, M., Li, Y., Hutchcroft, W., Velipasalar, S., and Kang, S. B. (2024). BGDNet: Background-guided Indoor Panorama Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1272-1281)
[37] Yun, Ilwi, et al. "Egformer: Equirectangular geometry-biased transformer for 360 depth estimation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
[38] Armeni, I., et al. "Joint 2d-3d-semantic data for indoor scene understanding. arXiv 2017." arXiv preprint arXiv:1702.01105 5.6 (2017).

Qingling Chang received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2015. She is a Master's Supervisor and an Associate Professor with Wuyi University and the Subdecanal of China–German Artificial Intelligence Institute. She has published more than ten papers included in SCI/EI. Her research interests include artificial intelligence, computer vision, and knowledge graph.

Jingheng Xu received the bachelor's degree in computer science and technology from Wuyi University, in 2024. His current research interest includes depth estimate and semantic segmentation.

Yan Cui is a Professor with the Faculty of Computer Science, Wuyi University, where he is also the Dean of the School of Intelligent Manufacturing. He is the Dean of China–Germany Artificial Intelligence Institute. His research interests include computer vision and computer graphic.

Yikui Zhai (Senior Member, IEEE) received his Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in June 2013. Since October 2007, he has been working with Wuyi University, Jiangmen, China, where he is a Full Professor now. He is also Associate Dean of the School of Electronics and Information Engineering in Wuyi University, since 2021. He has been a Visiting Scholar with Department of Computer Science, the Universita degli Studi di Milano, Italy, during June 2016 to June 2017, August 2023 and January 2024. His research interests include: Image Processing, Deep Learning, Optical Character Recognition, Object Detection, UAV Change Detection, Self Supervise Learning.

Pasquale Coscia (Senior Member, IEEE) received the Ph.D. degree in Industrial and Information Engineering from the Universita degli Studi della Campania Luigi Vanvitelli, Italy, in 2019. From 2019 to 2022, he was a post-doctoral researcher at the Universita degli Studi di Padova, Italy. He is currently an Assistant Professor at the Department of Computer Science of the Universita degli Studi di Milano, Italy. He is Co-chair of the Intelligent Measurement Systems Technical Committee (TC-22) of the IEEE Instrumentation and Measurement Society since 2023. His research activities are focused on theoretical, methodological, and applied aspects of computational intelligence for signal and image processing.

Angelo Genovese (Senior Member, IEEE) received the Ph.D. degree in computer science from the Universita degli Studi di Milan, Crema, Italy, in 2014. He has been a postdoctoral Research Fellow in computer science with the Universita degli Studi di Milan since 2014. He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. Original results have been published in over 30 papers in international journals, proceedings of international conferences, books, and book chapters. His current research interests include signal and image processing, three-dimensional reconstruction, computational intelligence technologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems.

Vincenzo Piuri (Fellow, IEEE) received the M.S. and Ph.D. degrees in computer engineering from Politecnico di Milan, Milan, Italy, in 1984 and 1988, respectively. He was the Department Chair with the University of Milan, Milan, from 2007 to 2012, where he has been a Full Professor since 2000. He was an Associate Professor with Politecnico di Milan from 1992 to 2000, a Visiting Professor with The University of Texas at Austin, Austin, TX, USA from 1996 to 1999, and a Visiting Researcher with George Mason University, Fairfax, VA, USA, from 2012 to 2016. He founded a startup company, Sensuresrl, Bergamo, Italy, in the area of intelligent systems for industrial applications (leading it from 2007 to 2010) and was active in industrial research projects with several companies. His main research and industrial application interests are intelligent systems, computational intelligence, pattern analysis and recognition, machine learning, signal and image processing, biometrics, intelligent measurement systems, industrial applications, distributed processing systems, Internet-of-Things, cloud computing, fault tolerance, application-specifi digital processing architectures, and arithmetic architectures. Dr. Piuri is an ACM Fellow.

Fabio Scotti (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003. He was an Assistant Professor with the Department of Information Technologies, Universita degli Studi di Milan, Milan, from 2002 to 2015, where he was an Associate Professor with the Department of Computer Science from 2015 to 2020. He has been a Full Professor with the Universita degli Studi di Milan since 2020. Original results have been published in over 150 papers in international journals, proceedings of international conferences, books, book chapters, and patents. His current research interests include biometric systems, machine learning and computational intelligence, signal and image processing, theory and applications of neural networks, 3-D reconstruction, industrial applications, intelligent measurement systems, and high-level system design. Prof. Scotti is an Associate Editor of the IEEE Transactions on Human–Machine Systems and the IEEE Open Journal of Signal Processing. He is serving as a Book Editor (Area Editor, section Less-Constrained Biometrics) of the Encyclopedia of Cryptography, Security, and Privacy (3rd Edition, Springer). He has been an Associate Editor of the IEEE Transactions on Information Forensics and Security, Soft Computing (Springer) and a Guest Coeditor of the IEEE Transactions on Instrumentation and Measurement.