360BEV: Panoramic Semantic Mapping for Indoor Bird's-Eye View

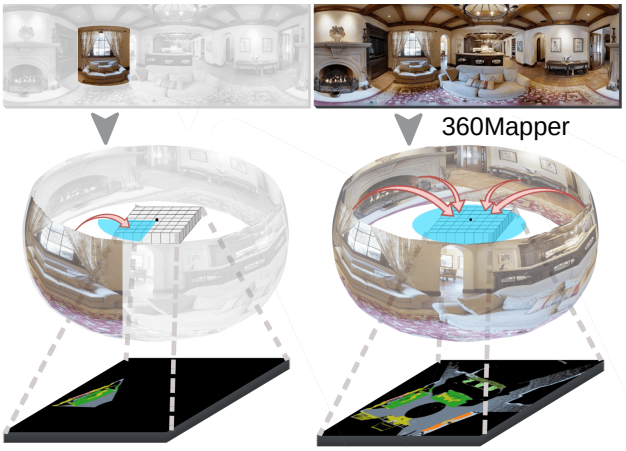
Zhifeng Teng[1,*], Jiaming Zhang[1,*,†], Kailun Yang[2], Kunyu Peng[1],
Hao Shi[3], Simon Reiß[1], Ke Cao[1], Rainer Stiefelhagen[1]
[1]Karlsruhe Institute of Technology, [2]Hunan University, [3]Zhejiang University

Abstract

*Seeing only a tiny part of the whole is not knowing the full circumstance. Bird's-eye-view (BEV) perception, a process of obtaining allocentric maps from egocentric views, is restricted when using a narrow Field of View (FoV) alone. In this work, mapping from 360° panoramas to BEV semantics, the **360BEV** task, is established for the first time to achieve holistic representations of indoor scenes in a top-down view. Instead of relying on narrow-FoV image sequences, a panoramic image with depth information is sufficient to generate a holistic BEV semantic map. To benchmark 360BEV, we present two indoor datasets, 360BEV-Matterport and 360BEV-Stanford, both of which include egocentric panoramic images and semantic segmentation labels, as well as allocentric semantic maps. Besides delving deep into different mapping paradigms, we propose a dedicated solution for panoramic semantic mapping, namely **360Mapper**. Through extensive experiments, our methods achieve 44.32% and 45.78% mIoU on both datasets respectively, surpassing previous counterparts with gains of +7.60% and +9.70% in mIoU.*[1]

1. Introduction

Semantic scene understanding has achieved remarkable performance on indoor- and outdoor scenes via pixel-wise semantic segmentation [22]. It can be utilized directly on a wide range of downstream applications, such as autonomous driving [10, 13], navigation in robotics [4, 6] or in assistive technologies [37] to name a few. Recently, Bird's-Eye-View (BEV) semantic perception [17] can be a solution for enabling a straightforward understanding of the environment and objects therein. While BEV semantic segmentation has gained traction in outdoor scenes for autonomous driving [17], BEV perception has not yet been extensively explored for indoor scenes, which are often characterized by complex and varied structures, objects, and challenging lighting conditions. For semantically mapping these indoor scenes, sequence-based methods [4, 6] were proposed, which have to process whole videos and entail a moving camera. As shown in Fig. 1a, (1) these methods rely on computationally expensive processing of entire sequences of video-frames due to the narrow Field of View of the pinhole camera, and (2) they are constrained to explore indoor mapping on synthetic simulators [26, 32], due to the lack of real indoor datasets. These drawbacks limit their applicability to real-world indoor semantic mapping.



(a) Narrow BEV (b) 360BEV

Figure 1. Semantic mapping from egocentric front-view images to allocentric BEV semantics. While (a) the narrow-BEV method has limited perception and map range, (b) 360BEV has an omnidirectional Field of View, yielding a more complete BEV map by using our 360Mapper model.

To solve these limitations, in this work, we introduce **360BEV** to achieve panoramic semantic mapping for indoor BEV, which is illustrated in Fig. 1b. Our considerations are twofold: (1) To unleash the potential of indoor semantic mapping in real-world scenarios, real indoor databases with BEV semantic labels are crucial; (2) To reduce the computational complexity of narrow-FoV sequence methods [4] (≥20 video-frames to process) or the complexity of multi-camera setups [17] (≥6 camera views needed), we leverage a single-frame 360° image with depth information and thus bypass multi-sensor calibration, syn-

*Equal contribution.
†Corresponding author (e-mail: jiaming.zhang@kit.edu).
[1]The datasets and code are available at the project page 360BEV.

chronization, and data fusion procedures. With this in mind and to enable 360BEV segmentation we present two real indoor BEV datasets, which are extended from the Matterport3D [5] and Stanford2D3D [3] datasets. First, the Front-View images captured by pinhole cameras from Matterport3D are extended to 360° panoramas for benchmarking on **360FV-Matterport**. Furthermore, for the first time, two BEV datasets, **360BEV-Matterport** and **360BEV-Stanford** are established to enable bird's-eye view panoramic semantic mapping, *i.e*., predicting a complete BEV semantic map from a single-frame 360° image with depth. Moreover, by decoupling the computationally expensive processing of sequences or multiple views, our direct 360BEV semantic mapping is more streamlined for generating indoor semantic maps.

However, spatial distortions and object deformations in panoramic images [38] severely harm the performance of methods proposed for narrow-range image [12,33] or multi-view perception [17]. Thus, to comprehensively investigate the established 360BEV task, we first revisit three possible projection paradigms, including: (1) *Early projection*, (2) *Late projection*, and (3) *Intermediate projection*. Based on our observation that intermediate features maintain dense information, we explore the intermediate projection paradigm and propose a dedicated solution for 360BEV mapping, which we call **360Mapper**. The challenge in this scheme resides in the feature conversion. While the prior BEVFormer [17] relied on multi-view perception and SM-Net [4] projects the extracted feature directly via the depth-based transformation index, which is not appropriate for panoramic imagery due to its distortions and deformations, we propose a new transformation method, the **Inverse Radial Projection (IRP)**, to project features from 2D to 3D representations using only depth information. An additional benefit is that the depth information helps maintain object shape and space layout after being transferred to top-down views, rendering the 2D reference index for the feature map as well as the BEV representation more accurate and consistent. Besides, unlike the deformable attention [17, 44] using multi-scale layers and fusion from multi-view cameras, we adopt **360Attention** with adaptive sampling offsets to extract information from omnidirectional feature maps, yielding the bird's-eye-view feature with less distortion in an adaptive manner. These are combined with the 2D index obtained by IRP to include a deformation-aware mechanism in 360 scenes, which in turn serves to compensate for the adverse effects of distortion. With these designs, our 360Mapper model represents a step towards a more complete and accurate indoor semantic mapping, which has important implications for downstream applications such as indoor navigation and scene understanding.

Through extensive experiments, the new 360BEV task is thoroughly benchmarked with two real indoor BEV datasets, three projection paradigms, and more than ten methods, respectively. Compared to the semantic mapping counterparts, our 360Mapper models achieve state-of-the-art performance, with mean intersection-over-union (mIoU) gains of $>7\%$ on the 360BEV-Matterport dataset and $>9\%$ on the 360BEV-Stanford dataset.

To summarize, we present the following contributions:

- A new *360BEV* task is introduced for the first time to address indoor semantic mapping via a single-frame panoramic image, decoupling complex processing of multi-view or sequence inputs.
- Two indoor BEV datasets, *i.e*., *360BEV-Matterport* and *360BEV-Stanford*, are extended with front-view panoramic images and BEV semantic labels, thoroughly benchmarking panoramic semantic mapping.
- *360Mapper* model – addressing spatial distortions and object deformations in panoramas – is proposed as a dedicated solution for interior panoramic semantic mapping and achieves state-of-the-art performance.

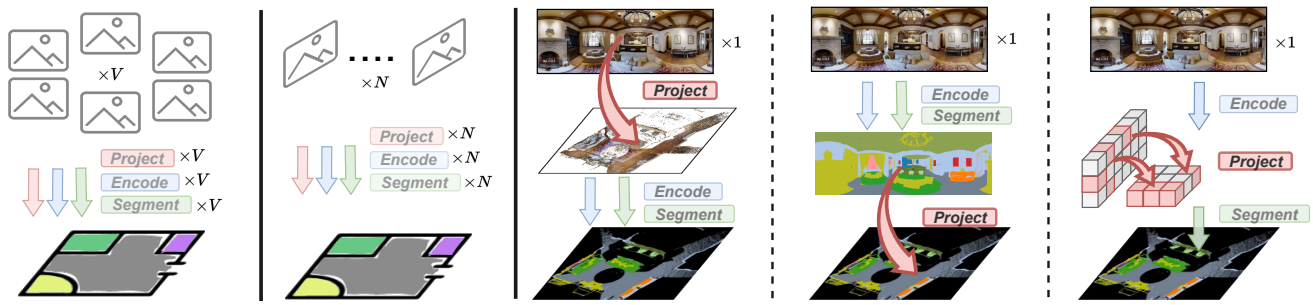
2. Related Work

2.1. Panoramic Semantic Segmentation

Image semantic segmentation [31, 33, 35, 40] has achieved great progress. In contrast to narrow-FoV perception, panoramic semantic segmentation [7–9, 14, 29, 30, 36, 38], yielding holistic scene understanding by using a single 360° front-view image, has received increasing attention in recent years. Besides, 3D60 [45] and Pano3D [1] datasets are generated for depth estimation from 360° images, but lack semantic labels. In indoor panorama segmentation, there are some benchmarks that provide synthetic [16, 41] and real [3] panoramic images and labels for training. Matterport3D [5] has large-scale panoramic images collected from 90 indoor buildings, yet, it has not been benchmarked due to the lack of corresponding panoramic semantic labels. To enable this, we generate the panoramic semantic segmentation labels by combining the original 18 pinhole camera labels regarding their camera transformation matrices. Therefore, a 360° Front-View (FV) dataset, *360FV-Matterport*, with large-scale real indoor scenes, is provided to facilitate panoramic semantic segmentation. Besides, the 360FV-Matterport dataset is required to perform the late-projection paradigm of BEV semantic mapping.

2.2. BEV Semantic Mapping

Apart from front-view image semantic segmentation, some previous work explored top-view semantic segmentation, known as semantic mapping [6, 22] in indoor scenes and bird's-eye-view semantic segmentation [17, 24] in outdoor driving scenes. The indoor semantic mapping methods can be divided into three categories according to the level of projection from the front view to the top-down



(a) Multi-view (b) Sequence-based (c) Early-projection (d) Late-projection (e) Intermediate-projection

Figure 2. **Paradigms of semantic mapping.** While the narrow-FoV (a) multi-view and (b) sequence-based methods rely on $V \geq 6$ and $N \geq 20$ views, the 360°-BEV (c) Early-, (d) Late-, and (e) Intermediate-projection methods use a single panorama.

view: *Early-projection* approaches [20, 28] are performed via general semantic segmentation methods, which first construct the BEV views from perspective images and then apply segmentation. Unfortunately, these pipelines lose fine-grained visual cues during the projection and thus result in unsatisfactory performance for small object segmentation. *Intermediate-projection* methods [4, 6] directly take front views as input for holistic indoor scene understanding, however, they work on synthetic data generated from Gibson [32] or Habitat [26] simulators and rely on time-consuming image sequences. For example, SMNet [4] gradually captures an average of 2,500 view-points for each floor to generate a semantic map for indoor scenes. Instead, we explore achieving efficient allocentric scene understanding via a single panorama image. The *Late-projection* pipeline [2, 11, 21, 25, 27] performs egocentric semantic segmentation and project labels to top-down views, which are sensitive to depth map and agent pose information, inevitably facing the projection error and under-fitting of model training, thus remaining a suboptimal solution. There are some BEV-related methods that leverage multiple perspective view sensors or LiDAR sensors and focus on outdoor object detection [17, 18, 34], optical flow estimation [15, 19], and semantic segmentation [23, 43]. Different from previous methods, our 360Mapper is carefully designed for learning indoor holistic representations by forwarding a single panorama without using multi-view images, image sequences, or point clouds.

3. Panorama Semantic Mapping (360BEV)

To investigate the 360BEV task, we analyze potential panoramic projection paradigms in Sec. 3.1. The generation and data statistics of the dataset are detailed in Sec. 3.2. To tackle the challenging panoramic semantic mapping, in Sec. 3.3 we present our solution **360Mapper** with the **Inverse Radial Projection** method and **360Attention** module, which enable distortion-aware feature processing.

3.1. 360 Projection Paradigms

As shown in Fig. 2, unlike multi-view methods relying on more than six views (V in Fig. 2a) and sequence-based methods using more than 20 narrow views (N in Fig. 2b), panoramic semantic mapping uses a single image with depth. We investigate three projection paradigms, *i.e.*, *how to process data from front-view panoramas to bird's-eye-view semantics*, which are:

(1) *Early projection:* ***Proj.***$\rightarrow$*Enc.*$\rightarrow$*Seg.* in Fig. 2c.
(2) *Late projection: Enc.*$\rightarrow$*Seg.*$\rightarrow$***Proj.*** in Fig. 2d.
(3) *Intermediate projection: Enc.*$\rightarrow$***Proj.***$\rightarrow$*Seg.* in Fig. 2e.

Based on these properties, we mainly explore 360BEV with intermediate projections, in which we identify the following challenges: In the feature extraction stage, spatial distortions and object deformations severely hinder the encoder from extracting representative features from the front-view panoramic image. For the intermediate feature projection, only depth information is utilized for consistent view transformation of high-dimensional features. In addition, many large objects in the front view (*e.g.*, *walls*) are projected to thin objects in the top-down view, which greatly impedes capturing wide-range features during projection.

3.2. 360FV and 360BEV Data Generation

360FV-Matterport. The original Matterport3D [5] was collected via narrow-FoV cameras. As shown in Fig. 3, we convert the 18 narrow-view images and annotations into the 360° format by using rotation-translation matrices.

360BEV-Stanford. The Stanford2D3D dataset [3] has front-view panoramic images and semantic labels. However, it lacks BEV semantic labels. As presented in Fig. 4, we utilize the spatial semantic information from the global XYZ image to generate the corresponding BEV semantic map. By applying orthographic projection, we generate the BEV semantic maps within a visible range as BEV ground truth, enabling end-to-end training from font-view images to top-down semantics.

360BEV-Matterport. Inspired by the global XYZ modal-

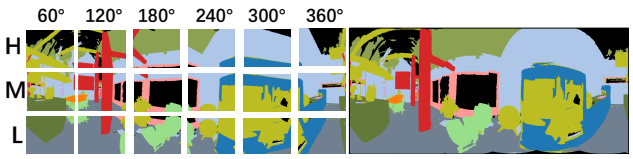


Figure 3. **360FV semantics generation** from 18 narrow views to a panoramic view on the 360FV-Matterport dataset. H, M, and L represent high, medium, and low positions, respectively.

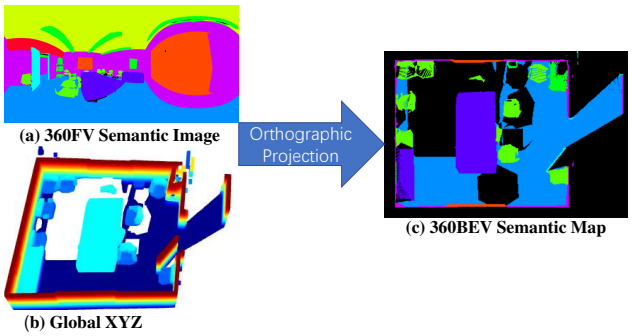


Figure 4. **360BEV semantics generation** by orthographic projection, from (a) the front-view semantic image and (b) the global XYZ image, to (c) the 360BEV semantic map.

ity [3], we generate a global XYZ for each panoramic image by using the provided depth ground truth. In order to generate BEV semantic ground truth corresponding to the panoramic view, several key steps must be considered. Firstly, a panoramic image can be processed as a sphere with rays shooting from the center of the sphere, where the camera is located.

$$\begin{aligned} \Theta_{i,j} &= \frac{i\pi}{H} + \frac{\pi}{2H}, \\ & i = \{0,\dots,H-1\},\ j = \{0,\dots,W-1\}, \\ \Phi_{i,j} &= -\frac{2\pi j}{W} + \pi - \frac{\pi}{W}, \\ & i = \{0,\dots,H-1\},\ j = \{0,\dots,W-1\}. \end{aligned} \tag{1}$$

Here, Θ and Φ are angle matrices of panoramic images with size $H \times W$, which consist of two dimensional Euler angular equivariant series. Given the representation in spherical coordinate systems, each 3D point $(X_{i,j}, Y_{i,j}, Z_{i,j})$ in the camera coordinate system will be obtained through the calculation in Eq. (2),

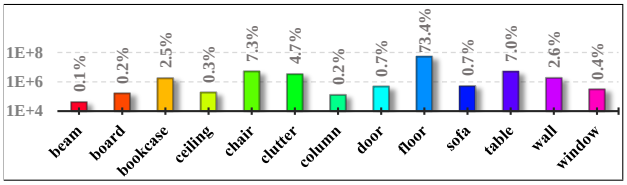
$$\begin{aligned} X_{i,j} &= D_{i,j} \cdot \sin(\Theta_{i,j}) \cdot \sin(\Phi_{i,j}), \\ Y_{i,j} &= D_{i,j} \cdot \cos(\Theta_{i,j}), \\ Z_{i,j} &= D_{i,j} \cdot \sin(\Theta_{i,j}) \cdot \cos(\Phi_{i,j}), \end{aligned} \tag{2}$$

where D is the panoramic depth information. After obtaining 3D points, the orthographic projection matrix P_v is applied to transform 3D coordinates to 2D panoramic BEV indices (u, v), which is presented in Eq. (3), where $[\mathbf{R}|\mathbf{t}]$ is the transformation matrix.

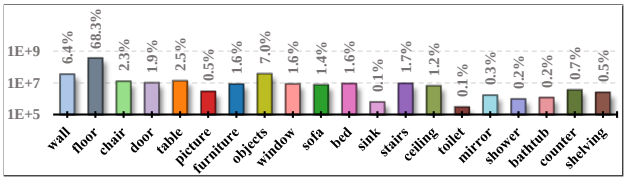
$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{R}^{-1} \begin{bmatrix} X_{i,j} \\ Y_{i,j} \\ Z_{i,j} \end{bmatrix} - \mathbf{t},$$
$$\underbrace{\begin{bmatrix} u \\ v \\ 0 \\ 1 \end{bmatrix} = P_v \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}}_{\text{Orthographic projection}}. \tag{3}$$

Table 1. **The data statistics** of the generated 360BEV-Matterport and 360BEV-Stanford datasets.

Dataset	#Scene	#Room	#Frame	#Category
`train`	5	215	1,040	13
`val`	1	55	373	13
360BEV-Stanford	6	270	1,413	13
`train`	61	–	7,829	20
`val`	7	–	772	20
`test`	18	–	2,014	20
360BEV-Matterport	86	2,030	10,615	20



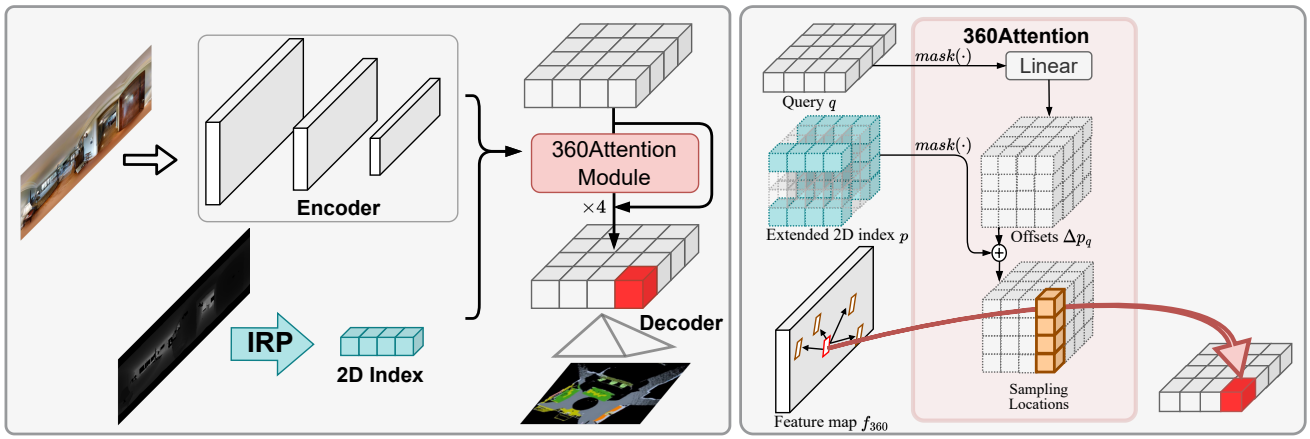
(a) Class distribution of 360BEV-Stanford dataset



(b) Class distribution of 360BEV-Matterport dataset

Figure 5. **Per-class pixel number (logarithmic) and frequency (%) distribution** of two 360BEV datasets.

Dataset statistics. As a result, two BEV datasets for panoramic semantic mapping are obtained. The detailed data statistics of 360BEV-Stanford and 360BEV-Matterport datasets are shown in Table 1. While the 360BEV-Stanford dataset has 13 classes and 1,413 images, the 360BEV-Matterport dataset includes 20 classes and 10,615 samples. In the Matterport3D dataset [5], there are 40 object categories in the dense annotation. However, many of them are relatively rare in the original dataset, *e.g.*, *TV* and *beam* ($\ll 0.1\%$), which are excluded. Thus, 360BEV-Matterport maintains the 20 most common object categories and merges some uncommon classes. Besides, we further present the per-class pixel number and per-class frequency in Fig. 5. It is worth noting that the $floor$ class has a much higher frequency on both datasets. This category is important for tasks that rely on complete maps, such as indoor navigation and is therefore also retained.



(a) 360Mapper model (b) 360Attention module

Figure 6. **Architecture of 360Mapper and the 360Attention module.** The 360Mapper model includes the encoder for extracting features from the front-view panoramic image, the 360Attention module for feature projection, and the decoder for parsing the projected feature to the BEV semantic map. The offsets are obtained by a linear layer and added with the 2D index that is obtained by Inverse Radial Projection (IRP), yielding the sampling locations for 360BEV feature projection.

3.3. Proposed Model: 360Mapper

Overall Architecture. As shown in Fig. 6, our end-to-end 360Mapper framework includes four steps: (1) The transformer-based backbone extracts features from the panoramic image. (2) The **Inverse Radial Projection (IRP)** module obtains a 2D index by projecting reference points generated by depth. (3) The **360Attention** module enhances the front-view feature by 2D index and generates offsets from BEV queries to eliminate the effects of distortion. (4) The lightweight decoder parses the projected feature map and predicts the semantic BEV map.

Inverse Radial Projection. Next, we propose a flexible projection method, the Inverse Radial Projection (IRP), for which the input of panoramic depth is included. We can easily obtain a top-view mask map by projecting from depth information. This mask map is then used to generate 3D reference points with the corresponding map height. 3D reference points are projected onto the sphere to generate 2D reference indexes, as shown in Eq. (4), where ID_h and ID_w represent the index values of the 2D reference for the height and width of the feature map, respectively. The 2D reference indexes are then used to locate the corresponding feature points on the encoded front-view feature map.

$$\begin{aligned} \Phi &= \tan^{-1}\frac{y}{x}, \\ \Theta &= \tan^{-1}\left(\frac{x}{z}\cdot\frac{1}{\cos(\Phi)}\right), \\ ID_h &= \left\lceil \frac{H\Theta}{\pi} \right\rceil, \\ ID_w &= \left\lceil \left(\frac{\Phi}{\pi}-\frac{1}{W}\right)\cdot\frac{W}{2} \right\rceil. \end{aligned} \tag{4}$$

Due to the distortions in the stitching process of the panorama, it is hard to project the 3D reference points exactly onto the 2D front-view plane by rotation and translation. Thus, we use the depth map to generate a map mask that better describes the shape of the map, so that the accurate projection with the mask not only makes the amount of data entering the 360Attention much smaller, which is conducive to the fast convergence of the model but also facilitates the use of sampling offsets for 360Attention.

360Attention. In Fig. 6b, the proposed 360Attention generates sampling offsets through the linear layer in an adaptive manner. Given the BEV query $\boldsymbol{q} \in \mathbb{R}^{N\times C_{Emb}}$ as input, where $N{=}h{\times}w$ is the length of query, a $mask(\cdot)$ operation is applied on $\boldsymbol{q}$ and $\boldsymbol{p}$ to mask out irrelevant points and 2D indexes according to the mask map M_{map} from IRP, which is crucial to keep $\boldsymbol{q}$ and $\boldsymbol{p}$ efficient and reducing computation of 360Attention ($\sum M_{map}{<}N$). The sampling offset $\Delta\boldsymbol{p}_{q,ij}$ and attention weight $\mathcal{A}_{ij}{\in}[0,1]$ are predicted through BEV query by linear layers respectively. The adaptive sampling offsets are then added to the extended 2D index $\boldsymbol{p}$ to obtain distortion-aware sampling locations. The 360Attention module can be denoted as:

$$\begin{aligned} &\text{360Attn}(\boldsymbol{q},\boldsymbol{p},\boldsymbol{f}_{360}) = \\ &\sum_{i=1}^{N_{\text{head}}}\mathcal{W}_i\sum_{j=1}^{N_{\text{point}}}\mathcal{A}_{ij}{\cdot}\boldsymbol{f}_{360}\left(mask\left(\boldsymbol{p}\right)+\Delta\boldsymbol{p}_{q,ij}\right), \end{aligned} \tag{5}$$

where $\boldsymbol{q}$, $\boldsymbol{p}$, and $\boldsymbol{f}_{360}$ indicate the query, the extended 2D index, and panoramic feature map, respectively. The linear layer $\mathcal{W}_i{\in}\mathbb{R}^{C\times(C/N_{head})}$ is specific to each attention head i, where C is the feature dimension and N_{head} is the number of heads. The attention weight $\mathcal{A}_{ij}$ represents the importance of the sampled points j, where $\sum\mathcal{A}_{ij}{=}1$.

Table 2. **Panoramic semantic segmentation (360FV)** on the Stanford2D3D dataset.

Method	Backbone	mIoU(%)
Tangent [8]	ResNet-101	45.6
SegFormer [33]	MiT-B2	51.9
HoHoNet [29]	ResNet-101	52.0
Trans4PASS [38]	MiT-B2	52.1
CBFC [42]	ResNet-101	52.2
Ours	MiT-B2	**54.3**

Table 3. **Panoramic semantic segmentation (360FV)** on the `val` set of 360FV-Matterport dataset.

Method	Backbone	mIoU(%)
HoHoNet [29]	ResNet-101	44.10
Trans4PASS [38]	MiT-B2	41.91
Trans4PASS+ [39]	MiT-B2	42.60
SegFormer [33]	MiT-B2	45.53
Ours	MiT-B2	**46.35**

The panoramic features $\boldsymbol{f}_{360}$ and the adaptive sampling locations $(mask(\boldsymbol{p})+\Delta\boldsymbol{p}_{q,ij})$ are aggregated using attention weights $\mathcal{A}_{ij}$ to produce a BEV output. Afterwards, the mask map M_{map} is applied to assemble the BEV output as $\boldsymbol{q}'\in\mathbb{R}^{N\times C_{Emb}}$. After being added with a residual term of $\boldsymbol{q}$, the BEV result from $\boldsymbol{q}+\boldsymbol{q}'$ is forwarded to the next 360Attention module.

Compared to the Spatial Cross-Attention module in BEVFormer [17], the difference lies in (1) Instead of relying on multi-view features across multiple cameras, our 360Attention module is designed to directly adopt adaptive sampling offsets to extract features from a single panoramic feature map. (2) Our module gets rid of the projection of 3D reference points to different image views using the projection matrix, thus compensating for the lack of front-view perception. (3) The mask operation is applied to maintain the BEV query efficient and adaptive to front-view panoramic features by using depth information as a bridge. Through these non-trivial designs, the BEV feature map generated by 360Attention is able to effectively neutralize the effects of front-view distortion.

4. Experiments

4.1. Implementation Details

We train 360Mapper models with 4 A100 GPUs with an initial learning rate of $6e^{-5}$, scheduled by the step strategy over 50 epochs. AdamW is the optimizer with epsilon $1e^{-8}$, weight decay is 0.01 and batch size is 4 on each GPU. The panoramic image size of 360FV-Matterport and Stanford2D3D [3] are both 512×1024. The resolution of panoramic images on both 360BEV-Stanford and 360BEV-Matterport datasets are 512×1024 as input for 360Mapper training, while the output BEV maps are set to 500×500, which correspond to a perception range of $10m\times10m$. Following [4, 6], evaluation metrics are pixel-wise accuracy (Acc), pixel recall (mRecall), precision (mPrecision), and mean Intersection-over-Union (mIoU).

4.2. Panorama Semantic Segmentation (360FV)

Results on Stanford2D3D. To verify the capacity to handle object deformations and image distortions, we first evaluate our method on front-view panoramic semantic segmentation. The results on the Stanford2D3D dataset are presented in Table 2. All results are averaged over 3 cross-validation folds. Thanks to the proposed 360Attention module, our 360Mapper model is better capable of handling deformations in panoramas, yielding 54.3% in mIoU, with >2% performance gains as compared to the previous state-of-the-art Trans4PASS [38] and CBFC [42]. The promising result in front-view panoramas has initially revealed the potential of our model in extracting 360° front-view features, which is crucial for the BEV semantic mapping task as well.

Results on 360FV-Matterport. For the first time, a large-scale 360FV-Matterport is brought to the community of front-view panoramic semantic segmentation. In Table 3, four state-of-the-art methods are selected and reproduced. Compared to the Trans4PASS [38] and Trans4PASS+ [39] models, our model has respective +4.44% and +3.75% improvements. Furthermore, our model surpasses RGB-D HoHoNet [29] and SegFormer [33] with +1.50% and +0.82% mIoU gains. The results indicate that our model can consistently achieve state-of-the-art performance on large-scale datasets for panoramic semantic segmentation.

4.3. Panorama Semantic Mapping (360BEV)

To thoroughly investigate the 360BEV task, we consistently analyze the early-, late-, and intermediate projections, as well as compare their state-of-the-art methods in both 360BEV benchmarks.

Results on 360BEV-Stanford. In Table 4, to study the Early projection mode, SegFormer [33] and SegNeXt [12] with different backbones, are selected, which merely reach unsatisfactory results. The results indicate that the pre-projected RGB maintains less rich spatial and visual information of front-view images. Using Late projection, SegFormer with the same MiT-B2 backbone achieves 18.65% mIoU and surpasses the one using Early projection, still yielding sub-optimal semantic mapping results. Interestingly, all methods using Intermediate projection obtain more than 30% mIoU. While using the same MiT-B2 backbone, our proposed 360Mapper achieves 45.78% with +9.70% gains compared to the baseline Trans4Map [6]. Further, our efficient model (MiT-B0) outperforms Trans4Map (MiT-B4) with +05.73% mIoU gains. With a stronger CNN backbone MSCA-B from SegNeXt [12], our method reaches the best score with 46.44%

Table 4. **Panoramic semantic mapping (360BEV)** on the 360BEV-Stanford dataset.

Method	Backbone	Acc	mRecall	mPrecision	mIoU
(1) Early projection: Proj.→Enc.→Seg.					
SegFormer [33]	MiT-B2	71.69	20.82	26.34	14.15
SegNeXt [12]	MSCA-B	79.77	34.13	47.39	25.85
(2) Late projection: Enc.→Seg.→Proj.					
HoHoNet [29]	ResNet101	70.01	31.62	30.46	18.49
Trans4PASS [38]	MiT-B2	65.73	31.08	33.15	17.86
Trans4PASS+ [39]	MiT-B2	66.11	38.06	34.14	20.44
SegFormer [33]	MiT-B2	70.50	30.97	30.65	18.65
(3) Intermediate projection: Enc.→Proj.→Seg.					
BEVFormer [17]	MiT-B2	85.50	40.22	51.71	31.69
Trans4Map [6]	MiT-B0	86.41	40.45	57.47	32.26
Trans4Map [6]	MiT-B2	86.53	45.28	62.61	36.08
Trans4Map [6]	MiT-B4	86.99	46.18	58.19	36.69
Ours	MiT-B0	92.07	50.14	65.37	42.42 (+10.16)
Ours	MiT-B2	**92.80**	53.56	67.72	45.78 (+09.70)
Ours	MSCA-B	92.67	**55.02**	**68.02**	**46.44**

Table 5. **Panoramic semantic mapping (360BEV)** on the *val* set of 360BEV-Matterport dataset.

Method	Backbone	Acc	mRecall	mPrecision	mIoU
(1) Early projection: Proj.→Enc.→Seg.					
SegFormer [33]	MiT-B2	68.12	41.33	45.25	29.22
SegNeXt [12]	MSCA-B	68.53	42.13	46.12	30.01
(2) Late projection: Enc.→Seg.→Proj.					
HoHoNet [29]	ResNet101	62.84	38.99	44.22	26.21
Trans4PASS [38]	MiT-B2	55.99	29.59	40.91	20.07
Trans4PASS+ [39]	MiT-B2	57.89	32.75	40.93	21.58
SegFormer [33]	MiT-B2	62.98	41.84	45.30	27.78
(3) Intermediate projection: Enc.→Proj.→Seg.					
BEVFormer [17]	MiT-B2	72.99	43.61	51.70	32.51
Trans4Map [6]	MiT-B0	70.19	44.31	50.39	31.92
Trans4Map [6]	MiT-B2	73.28	51.60	53.02	36.72
Trans4Map [6]	MiT-B4	73.51	50.78	56.67	38.04
Ours	MiT-B0	75.44	48.80	56.01	36.98 (+5.06)
Ours	MiT-B2	78.80	59.54	59.97	44.32 (+7.60)
Ours	MSCA-B	**78.93**	**60.51**	**62.83**	**46.31**

in mIoU, which indicates 360Mapper is flexible to both CNN- and Transformer-based backbones.

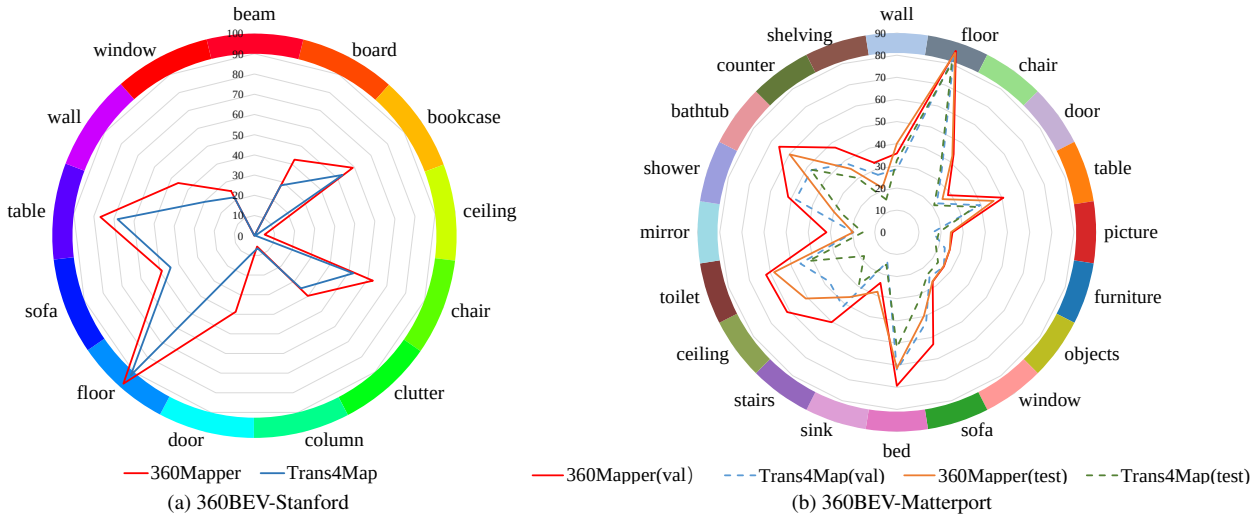
Results on 360BEV-Matterport. In Table 5, we further present the results on the 360BEV-Matterport dataset. SegFormer [33] and SegNeXt [12] adopt Early projection and show better performance than the Late projection ones. The reason for this is Late projection methods are constrained by their lower performance in front-view semantic segmentation, which affects the projected BEV semantic maps. In contrast, using Intermediate projection, our 360Mapper models based on two different model scales, *i.e.*, MiT-B0 and MiT-B2, show overall promising performance with 36.98% and 44.32% in mIoU, respectively. Compared to the previous state-of-the-art Trans4Map [6] (MiT-B2), our approach with MiT-B2 has improvements by +5.52% in accuracy, +7.94% in mRecall, +6.95% in mPrecision, and +7.60% in mIoU. Surprisingly, our 360Mapper with MiT-B2 outperforms Trans4Map with MiT-B4 with +6.28% in mIoU. Besides, to compare multi-view methods, we reproduce BEVFormer [17] by using a single panorama instead of six views of pinhole cameras. Our 360Mapper outperforms BEVFormer (MiT-B2) with +11.81% mIoU. Furthermore, we verify the flexibility of 360Mapper by using a CNN-based MSCA-B backbone [12], which obtains the highest mIoU score with 46.31%. All results are in line with our observation that Intermediate projection can preserve dense visual cues and long-range information from front-view panoramas, and deliver more valuable context for BEV semantic mapping, leading to this superiority of 360Mapper, as compared to the other paradigms.

Table 6. **Analysis of offset mechanisms in 360Attention and backbone variants** on 360BEV-Matterport dataset.

Methods	Backbone	#Param	FLOPs	mIoU
❶ Ours (360Attention offset)	MiT-B0	04.60M	248.57G	36.98
❷ Ours (360Attention offset)	MiT-B2	26.30M	283.94G	44.32
❸ Ours (360Attention offset)	MiT-B4	62.91M	341.34G	**45.53**
❹ Ours (Multi-scale offset)	MiT-B2	26.43M	284.17G	43.65 (-0.67)
❺ Ours (Fixed-range offset)	MiT-B2	26.30M	283.44G	43.28 (-1.04)
❻ Ours (Separate offset)	MiT-B2	26.19M	279.18G	42.82 (-1.50)
❼ Ours (360Attention offset)	MSCA-B	27.69M	274.59G	**46.31** (+1.99)

Per-class Results. To study the per-class performance on both 360BEV datasets, we present the comparison results in Fig. 7. For comparison, both the baseline Trans4Map and our 360Mapper model are based on the same backbone, *i.e.*, MiT-B2. On the 360BEV-Stanford dataset (Fig. 7a), our 360Mapper model has significant gains on most of categories, such as *board* (>14%), *wall* (> 16%), *door* (>28%), etc. On the 360BEV-Matterport dataset (Fig. 7b), it is readily apparent that our model can better recognize the *chairs* and *tables*, yielding >6% IoU gains compared to Trans4Map [6]. On the test set of the 360BEV-Matterport dataset, our 360Mapper obtains IoU gains with >12% and >15% on the *sink* and *toilet* classes, as compared to Trans4Map. Overall, the consistent improvements on both datasets show the superiority of our 360Mapper on panoramic semantic mapping.

Analysis of 360Attention. To better understand 360Attention, we further conduct an analysis of the offset mechanisms in 360Attention and the backbone selection, in Table 6. First, in ❶❷❸, we select three model scales, *i.e.*, MiT-B0, MiT-B2, and MiT-B4, to verify the effect of model capacity in 360Attention. The three models obtain good performance, showing that 360Attention has positive effects in different model scales. Besides, different offset schemes are compared among ❷❹❺❻, which are deformable, multi-scale, fixed-range, and separate offset. All of them have the same MiT-B2 backbone. Here, ❷ shows the superi-



(a) 360BEV-Stanford (b) 360BEV-Matterport

Figure 7. **Distribution of per-class semantic mapping results** (per-class IoU in %) on the 360BEV-Stanford and the 360BEV-Matterport datasets. Compared to the baseline model Trans4Map [6], our 360Mapper models achieve overall better 360BEV results.

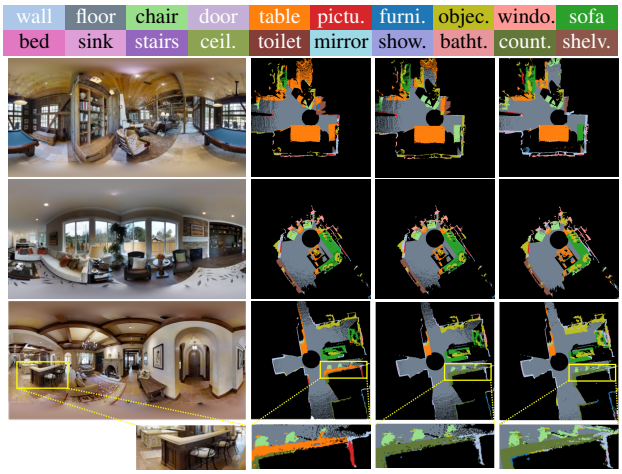


Figure 8. **Qualitative analysis** on the 360BEV-Matterport dataset. Black regions are `void`. Zoom in for a better view.

ority of deformable offset which has a better performance (44.32%). However, these comparable results prove that our 360Attention design is robust to offset mechanisms. Further, to analyze the effect of backbone selection, we choose transformer-based MiT-B2 [33] and CNN-based MSCA-B [12] as in ❷❼. A stronger backbone [12] shows a further improvement of mIoU (+1.99%), which shows the flexibility of our approach regarding the backbone variants.

4.4. Qualitative Analysis

To analyze the predicted semantic maps, we visualize the results from the validation set of the 360BEV-Matterport dataset. In Fig. 8, from left to right are input images, results of baseline [6], results of our 360Mapper, and ground truth. Thanks to the IRP projection and 360Attention, the segmentation results of 360Mapper are much better. In the first scene in Fig. 8, 360Mapper is able to successfully classify *chairs*, while the baseline model fails, predicting several *tables* and misclassifying the distant ground as another *table*. In the second scene, the segmentation of the *tables* derived by the baseline is incomplete. Furthermore, in the last zoomed-in scene, 360Mapper provides accurate semantic maps, such as in *counter*, *chair*, and *wall* categories, whereas the baseline Trans4Map [6] misclassifies them as *tables* and *doors*. Based on the qualitative analysis, our 360Mapper can effectively handle object deformations and image distortions, yielding better BEV semantic maps.

5. Conclusion

In this paper, we introduce 360BEV, a novel task to conduct panoramic semantic mapping in indoor environments, *i.e.*, from a single panoramic image to a holistic BEV semantic map. To enable this, we present 360BEV-Matterport and 360BEV-Stanford, extending off-the-shelf datasets for the presented 360BEV task. We revisit existing transformation paradigms and propose 360Mapper, a novel end-to-end architecture specifically designed for panoramic semantic mapping. As a consequence, 360Mapper outperforms state-of-the-art counterparts by clear margins.

Acknowledgement. This work was supported in part by the "KIT Future Fields" project, in part by the Ministry of Science, Research and the Arts of Baden-Wurttemberg (MWK) through the Cooperative Graduate School Accessibility through AI-based Assistive Technology (KATE) under Grant BW6-03, in part by the Federal Ministry of Education and Research (BMBF) through a fellowship within the IFI program of the German Academic Exchange Service (DAAD), and in part by Hangzhou SurImage Technology Company Ltd. We thank HoreKA@KIT, HAICORE@KIT, and bwHPC supercomputer partitions.

References

[1] Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiros Sterzentsenko, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. Pano3D: A holistic benchmark and a solid baseline for 360° depth estimation. In *CVPRW*, 2021. 2

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018. 3

[3] Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2D-3D-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 2, 3, 4, 6

[4] Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. Semantic MapNet: Building allocentric semantic maps and representations from egocentric views. In *AAAI*, 2021. 1, 2, 3, 6

[5] Angel Chang, Angela Dai, Thomas 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*, 2017. 2, 3, 4

[6] Chang Chen, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Trans4Map: Revisiting holistic bird's-eye-view mapping from egocentric images to allocentric semantics with vision transformers. In *WACV*, 2023. 1, 2, 3, 6, 7, 8

[7] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *ICML*, 2019. 2

[8] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *CVPR*, 2020. 2, 6

[9] Carlos Esteves, Ameesh Makadia, and Kostas Daniilidis. Spin-weighted spherical CNNs. In *NeurIPS*, 2020. 2

[10] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *T-ITS*, 2020. 1

[11] Margarita Grinvald, Fadri Furrer, Tonci Novkovic, Jen Jen Chung, Cesar Cadena, Roland Siegwart, and Juan Nieto. Volumetric instance-aware semantic mapping and 3D object discovery. *RA-L*, 2019. 3

[12] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. SegNeXt: Rethinking convolutional attention design for semantic segmentation. In *NeurIPS*, 2022. 2, 6, 7, 8

[13] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 2020. 1

[14] Chiyu Max Jiang, Jingwei Huang, Karthik Kashinath, Prabhat, Philip Marcus, and Matthias Nießner. Spherical CNNs on unstructured grids. In *ICLR*, 2019. 2

[15] Kuan-Hui Lee, Matthew Kliemann, Adrien Gaidon, Jie Li, Chao Fang, Sudeep Pillai, and Wolfram Burgard. PillarFlow: End-to-end birds-eye-view flow estimation for autonomous driving. In *IROS*, 2020. 3

[16] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. InteriorNet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *BMVC*, 2018. 2

[17] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 1, 2, 3, 6, 7

[18] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3D object detection. In *ECCV*, 2022. 3

[19] Chenxu Luo, Xiaodong Yang, and Alan Yuille. Self-supervised pillar motion learning for autonomous driving. In *CVPR*, 2021. 3

[20] Gellert Mattyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Enhancing road maps by parsing aerial images around the world. In *ICCV*, 2015. 3

[21] Daniel Maturana, Po-Wei Chou, Masashi Uenoyama, and Sebastian Scherer. Real-time semantic mapping for autonomous off-road navigation. In *FSR*, 2018. 3

[22] Branislav Mičušík and Jana Košecká. Semantic segmentation of street scenes by superpixel co-occurrence and 3D geometry. In *ICCVW*, 2009. 1, 2

[23] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *RA-L*, 2020. 3

[24] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. BEVSegFormer: Bird's eye view semantic segmentation from arbitrary camera rigs. In *WACV*, 2023. 2

[25] Teng Ran, Liang Yuan, Jianbo Zhang, Dingxin Tang, and Li He. RS-SLAM: A robust semantic SLAM in dynamic environments based on RGB-D sensor. *IEEE Sensors Journal*, 2021. 3

[26] Manolis Savva, Jitendra Malik, Devi Parikh, Dhruv Batra, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, and Vladlen Koltun. Habitat: A platform for embodied AI research. In *CVPR*, 2019. 1, 3

[27] Sunando Sengupta, Paul Sturgess, Lubor Ladicky, and Philip H. S. Torr. Automatic dense visual semantic mapping from street-level imagery. In *IROS*, 2012. 3

[28] Suriya Singh, Anil Batra, Guan Pang, Lorenzo Torresani, Saikat Basu, Manohar Paluri, and C. V. Jawahar. Self-supervised feature learning for semantic segmentation of overhead imagery. In *BMVC*, 2018. 3

[29] Cheng Sun, Min Sun, and Hwann-Tzong Chen. HoHoNet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, 2021. 2, 6, 7

[30] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *ECCV*, 2018. 2

[31] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2021. 2

[32] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *CVPR*, 2018. 1, 3

[33] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 2, 6, 7, 8

[34] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. BEVFormer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *CVPR*, 2023. 3

[35] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 2

[36] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. Orientation-aware semantic segmentation on icosahedron spheres. In *ICCV*, 2019. 2

[37] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen. Trans4Trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world. In *ICCVW*, 2021. 1

[38] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *CVPR*, 2022. 2, 6, 7

[39] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *arXiv preprint arXiv:2207.11860*, 2022. 6, 7

[40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[41] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large photo-realistic dataset for structured 3D modeling. In *ECCV*, 2020. 2

[42] Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Zhijie Shen, and Yao Zhao. Complementary bi-directional feature compression for indoor 360° semantic segmentation with self-distillation. In *WACV*, 2023. 6

[43] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *CVPR*, 2022. 3

[44] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2

[45] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, Federico Alvarez, and Petros Daras. Spherical view synthesis for self-supervised 360° depth estimation. In *3DV*, 2019. 2