Deep Single Image Camera Calibration by Heatmap Regression to Recover Fisheye Images Under Manhattan World Assumption Supplementary

Nobuhiko Wakai[1] Satoshi Sato[1] Yasunori Ishii[1] Takayoshi Yamashita[2]
[1] Panasonic Holdings Corporation [2] Chubu University
{wakai.nobuhiko,sato.satoshi,ishii.yasunori}@jp.panasonic.com takayoshi@isc.chubu.ac.jp

Structure of this paper. In this supplementary material, we present some details omitted from the main paper: the novelty of our method in Section S1, the limitations of our method associated with indoor scenes in Section S2, extended related work of panoramic images in Section S3, details of our method of rotation estimation and auxiliary diagonal points (ADPs) in Section S4, and additional experimental results of the vanishing point (VP) estimator and the whole of our method in Section S5.

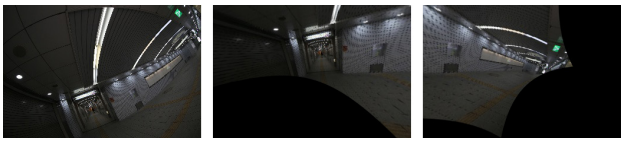
S1. Novelty

To describe the novelty of the paper, we again outline our major contributions:

1. We propose a heatmap-based VP estimator for recovering the rotation from a single image to achieve higher accuracy and robustness than geometry-based methods using arc detectors.
2. We introduce ADPs with an optimal 3D arrangement based on the spatial uniformity of regular octahedron groups to address the lack of VPs in an image.

We explain the novelty of the paper, along with our contributions, in the remainder of this section.

Heatmap-based vanishing point estimator. As the first contribution, our heatmap-based VP estimator achieved the detection of VPs and ADPs (VP/ADPs) in general scene images. By contrast, conventional geometry-based methods [1, 6, 12, 22] use arc detectors for estimating VPs. However, detection using arc detectors tends to fail in general scene images, such as images of trees lining a street. Furthermore, our VP estimator can robustly provide extrinsic camera parameters as VP/ADPs, in contrast to conventional learning-based methods [7, 17, 18] that use regressors without heatmaps. Our robust image-based method will contribute to subsequent studies; that is, robust camera rotation estimated by our method is useful for improving the performance of geometry-related tasks, such as simultaneous localization and mapping [5, 25].



Input Front direction Side direction

Figure S1. Qualitative results for indoor images from an off-the-shelf fisheye camera (ID 1). Far left image: input image. Right two images: the results of our method (HRNet-W32) trained using SL-MH (front and side directions).

Auxiliary diagonal points. As the second contribution, our proposed ADPs provide geometric cues that geometry-based methods cannot use; however, our heatmap-based VP estimator extracts these cues in general scene images. This approach of extracting geometric cues suggests that deep neural networks have the potential to obtain geometric cues that geometry-based methods cannot address. Similarly to our method, we believe that learning-based methods can use ADPs to improve their performance in geometry-related tasks, such as calibration, stereo matching, and simultaneous localization and mapping. In calibration, ADPs provide strong cues to compensate for the lack of VPs in images. Therefore, our method substantially outperformed both geometry-based [6, 12] and learning-based [7, 17, 18] state-of-the-art methods.

As described above, our major contributions have sufficient novelty to distinguish them from previous studies using both geometry-based and learning-based methods. Furthermore, we believe that our networks and ADPs will contribute to subsequent studies in many areas of computer vision, and are not limited to calibration.

S2. Limitations

Figure S1 shows the qualitative results of indoor scenes obtained by our method. We captured the input image using an off-the-shelf fisheye camera (ID 1 [18]) at an intersection in

an underpass. The indoor image degraded the performance of our method because of the domain gap between indoor and outdoor environments. In this paper, we focused on outdoor scenes following the studies of conventional learning-based methods [7, 17, 18]. We believe that subsequent studies will be able to extend this work to address the variety of indoor scenes.

S3. Extended related work

Due to the space limitations of the main paper, we review extended related work of panoramic images, such as equirectangular projection. These panoramic images are captured using panoramic cameras that are not fisheye cameras. However, both panoramic cameras and fisheye cameras have the same characteristics associated with large fields of view and distorted images. A typical task using panoramic images is panoramic depth estimation. In addition to depth estimation, panoramic depth completion is also described below.

Panoramic depth estimation. The task of the panoramic depth estimation is the estimation of dense depth maps from an RGB panoramic image. For an equirectangular projection, learning-based approaches can estimate dense depth maps from an image. An end-to-end depth estimation network was proposed by Wang *et al*. [19]. This neural network consists of two-branch neural networks processing the equirectangular projection and the cub-map projection with fusion blocks to leverage both projections. Eder *et al*. [3] proposed a tangent image spherical representation to alleviate the distortion of panoramic images. To improve accuracy and inference speed, Sun *et al*. [15] proposed a horizon-to-dense module relaxing the per-column output shape constraint. In addition to these convolutional neural networks, Shen *et al*. [13] proposed a Transformer-based method to improve accuracy. These panoramic depth estimation methods can only handle panoramic images in an equirectangular projection.

Panoramic depth completion. In contrast to panoramic depth estimation, panoramic depth completion is the estimation of dense depth maps from panoramic depth with missing areas. Yan *et al*. [23] proposed a pioneering method for the task of panoramic depth completion from a single 360° RGB-D pair. The multi-modal masked pre-training of this method generates shared random masks to make incomplete RGB-D pairs. This pre-training strategy allows networks to complete panoramic depth accurately. In addition, Yan *et al*. [24] also proposed a distortion-aware loss for the distortion of equirectangular projection and an uncertainty-aware loss for the inaccuracy in non-smooth regions. The proposed method using these loss functions achieved high accuracy for panoramic depth completion. These panoramic depth completion methods require RGB-D panoramic images.

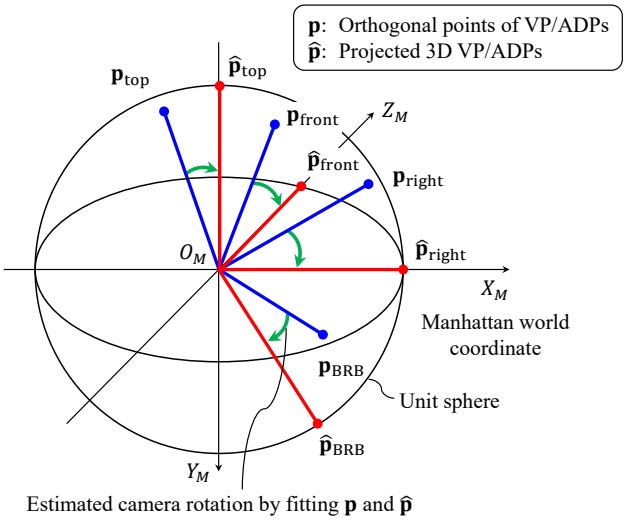


Figure S2. Projected 3D VP/ADPs and orthogonal points of VP/ADPs in the Manhattan world to estimate camera rotation. These orthogonal points are obtained as VP/ADPs without camera rotation; that is, pan, tilt, and roll angles are 0°. Four VP/ADPs of the labels in the front, right, top, and back-right-bottom (BRB) are shown in a unit sphere as an example of VP/ADPs.

As described above, these methods require panoramic input images captured using specific devices, that is, panoramic cameras. Additionally, panoramic images in the equirectangular projection are captured by upright or calibrated cameras. To satisfy these settings, we need to control the environments. Therefore, the networks for panoramic images cannot address deep single image camera calibration using fisheye images, which are rotated and distorted to varying degrees.

S4. Details of our method

In this section, we explain the details of our method of rotation estimation and describe the ADPs related to VPs.

S4.1. Rotation estimation

The details of rotation estimation are described in Section 3.3 (main paper). As described there, the estimation is performed by fitting two sets of world coordinates, which is known as the absolute orientation problem [21]. One set of world coordinates U consists of the 3D VP/ADPs, $\mathbf{p}$, projected by backprojection [18] using the camera parameters on condition that the rotation matrix $\mathbf{R}$ is a unit matrix and the translation vector $\mathbf{t}$ is a zero-vector. The other set $\hat{U}$ consists of the 3D points, $\hat{\mathbf{p}}$, that correspond to these VP/ADPs along the orthogonal Manhattan world coordinates, as shown in Figure S2. The absolute orientation problem is to fit the two sets, U and $\hat{U}$, by rotation, translation, and scaling. We focus only on rotation because VP/ADPs

are in a unit sphere.

It should be noted that this problem cannot be solved in the case of two or fewer VP/ADPs. To handle this condition, we add additional points using the cross-product operation. 1) In the case of two VP/ADPs, an additional point is calculated by the cross-product of the two position vectors of the VP/ADPs. 2) In the case of one VP/ADP, a temporal point on the unit sphere is added, whose direction is orthogonal to that of the VP/ADP. An additional point is calculated by the cross-product of the two position vectors of the temporal point and the VP/ADP. One of the angles (among the pan, tilt, and roll angles) of the temporal point is replaced by $0°$. 3) In the case of no VP/ADPs, $0°$ is used for the pan, tilt, and roll angles.

Conventional methods to solve the absolute orientation problem are based on singular value decomposition. To reduce the computational costs, the optimal linear attitude estimator [8, 10] was proposed, which uses skew-symmetric matrices instead of singular value decomposition. We describe the procedure for obtaining the pan, tilt, and roll angles because conventional calibration methods report results with these angles rather than the Rodrigues vector. First, we estimate the camera rotation as the Rodrigues vector, $\mathbf{g} = (g_x, g_y, g_z)^\mathsf{T}$, using this optimal linear attitude estimator. Second, we convert the Rodrigues vector $\mathbf{g}$ to an optimal quaternion, $\hat{q} = (\hat{q}_x, \hat{q}_y, \hat{q}_z, \hat{q}_w)$, using the equation

$$\hat{q} = \frac{q}{\sqrt{q^\mathsf{T} q}}, \tag{S1}$$

where $q = (g_x, g_y, g_z, 1)$ [10]. Third, we obtain a rotation matrix $\mathbf{R}$ from the quaternion $\hat{q}$ using the equation

$$\mathbf{R} = \begin{bmatrix} a_w + a_x - 1 & a_{xy} - a_{zw} & a_{xz} + a_{yw} \\ a_{xy} + a_{zw} & a_w + a_y - 1 & a_{yz} - a_{wx} \\ a_{xz} - a_{yw} & a_{yz} + a_{wx} & a_w + a_z - 1 \end{bmatrix}, \tag{S2}$$

where

$$\begin{aligned} (a_x, a_y, a_z, a_w) &= (2\hat{q}_x^2, 2\hat{q}_y^2, 2\hat{q}_z^2, 2\hat{q}_w^2), \\ (a_{xy}, a_{yz}, a_{zw}, a_{wx}) &= (2\hat{q}_x\hat{q}_y, 2\hat{q}_y\hat{q}_z, 2\hat{q}_z\hat{q}_w, 2\hat{q}_w\hat{q}_x), \\ (a_{xz}, a_{yw}) &= (2\hat{q}_x\hat{q}_z, 2\hat{q}_y\hat{q}_w). \end{aligned}$$

Finally, we calculate the pan, tilt, and roll angles by decomposing the rotation matrix. However, this decomposition is not unique without constraints. To solve this problem, we determined the pan, tilt, and roll angles for which the mean absolute angle errors between the estimated and ground-truth (GT) angles are the smallest, for both our method and conventional methods. It should be noted that, in our method, the estimated Rodrigues vector is directly used for applications, and the decomposition described above was employed to evaluate angle errors.

S4.2. Symmetry of auxiliary diagonal points

We describe the optimal arrangement of ADPs in detail. Our calibration method requires at least two unique axes to estimate camera rotation without ambiguity. It is possible to add VP-related points, such as ADPs; however, increasing the number of points causes unstable optimization. To address this trade-off, we analyze the arrangement of VP/ADPs with respect to 3D spatial uniformity and the number of points.

In world coordinates at a unit sphere, VPs form a regular octahedron, shown in Figure S3(a). This regular octahedron has the symmetry of the regular octahedron groups, whose rotational symmetry has six axes in C_2, four axes in C_3, and three axes in C_4. It should be noted that C_n represents the rotational symmetry using Schoenflies notation; that is, C_n is $(360°/n)$-rotational symmetry in Figure S3(b). We need to define VP-related points along C_2, C_3, or C_4 to maintain the symmetry of the regular octahedron groups; that is, these points are on the axes or form axial symmetry. Because of the trade-off described above, we focus on arrangements with a small number of points. Figure S3 shows arrangements of our proposed ADPs and candidate points, as explained below.

First, we explain the arrangement of ADPs illustrated in Figure S3(c). Along the C_3 axes, ADPs are located at the eight corners of a cube. The minimum angle formed by two axes, α, is $54.7°$. This angle expresses the magnitude of the 3D spatial uniformity; that is, biased arrangements decrease α. Second, C_3-based auxiliary points, of which there are 24, are defined along the C_3 axes (C_3-axial symmetry), as shown in Figure S3(d). The number of points (24) is the second smallest number of points for the C_3 axes because the C_3 axes have $120°$-rotational symmetry, yielding 24 points (8 axes $\times 360°/120°$-rotational symmetry). Third, C_4-based auxiliary points, of which there are 12, are defined along the C_4 axes (C_4-axial symmetry), as shown in Figure S3(e). Each point is located on the bisector of an angle between two orthogonal axes: two axes among X_M, Y_M, and Z_M. We use these axial-symmetric points because VPs are located along the C_4 axes. Fourth, C_2-based auxiliary points, of which there are 12, are defined along the C_2 axes, as shown in Figure S3(f). Each point is located in the middle of the edge of a cube. It should be noted that we can assume other arrangements satisfying the symmetry of the regular octahedron groups, in addition to those above. These other arrangements have more auxiliary points than those in (c), (d), (e), and (f) have for each symmetric axis. Therefore, we focus on the arrangements illustrated in Figure S3 because many auxiliary points cause unstable optimization in training.

Of all the cases discussed above, the minimum number of points is eight, as shown in Figure S3(c). The number of ADPs (8) is smaller than that of C_2-based and C_4-based

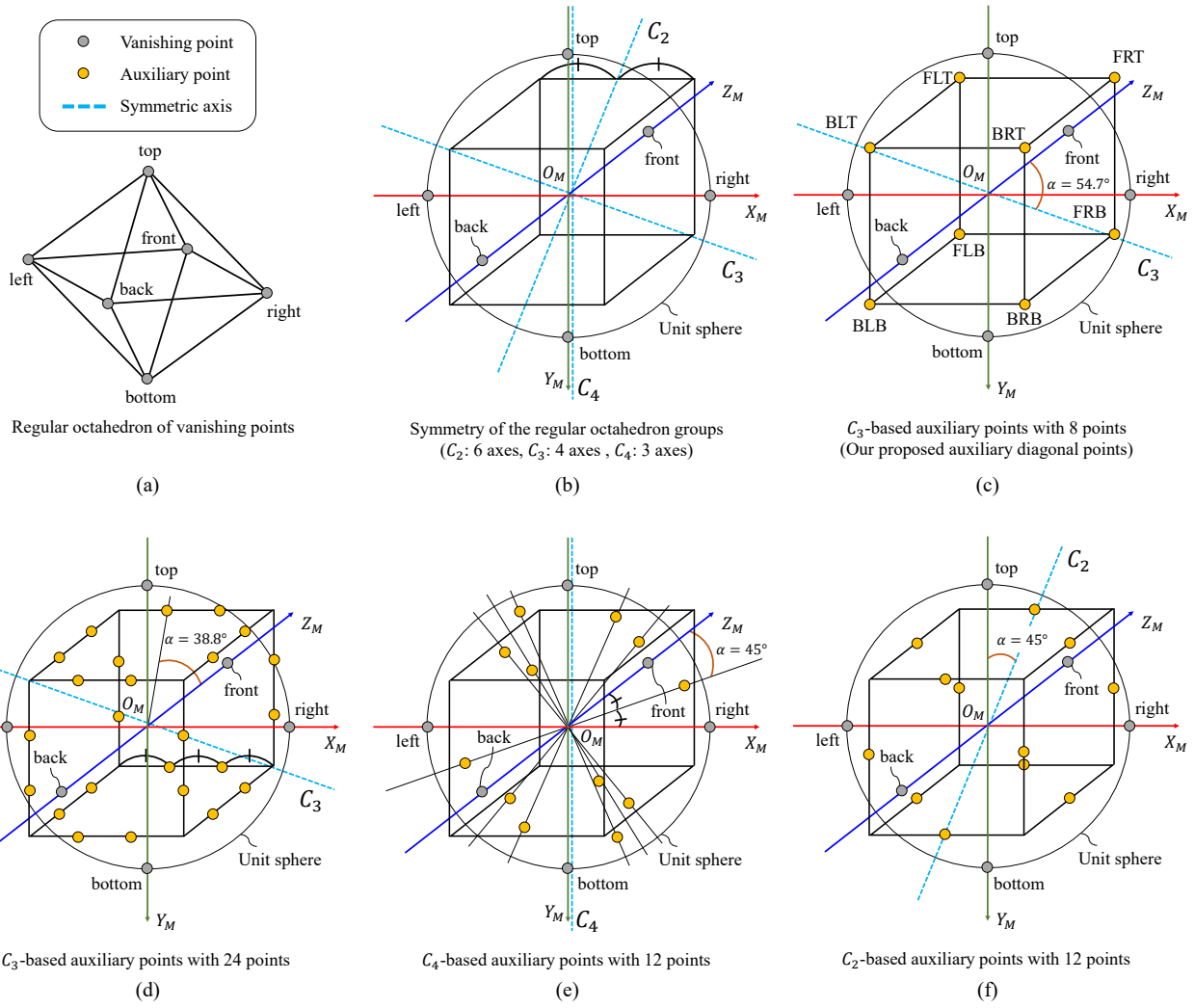


Figure S3. Arrangements of VPs and ADPs. (a) A regular octahedron formed by VPs. (b) Symmetry axes of C_2, C_3, and C_4 in the symmetry of the regular octahedron groups. (c) An arrangement of ADPs. (d) An arrangement of C_3-based auxiliary points with 24 points. (e) An arrangement of C_4-based auxiliary points with 12 points. (f) An arrangement of C_2-based auxiliary points with 12 points.

Table S1. Comparison of the numbers of auxiliary points and minimum axis angles

Arrangement		Number of auxiliary points	Minimum axis angle $\alpha \uparrow$
C_3-based auxiliary points in Figure S3(c)	(ADPs)	8	54.7°
C_3-based auxiliary points in Figure S3(d)		24	38.8°
C_4-based auxiliary points in Figure S3(e)		12	45.0°
C_2-based auxiliary points in Figure S3(f)		12	45.0°

auxiliary points (12). In addition, ADPs have 3D spatial uniformity with respect to the minimum axis angle α, as shown in Table S1. Therefore, we use ADPs, which have the optimal arrangement in the case of eight points, for our calibration method.

As described in Section 4.2 (main paper), our method was able to estimate a unique rotation for over 98% of the images in our experiments because of the arrangement of optimal 3D spatial uniformity, as presented in Table S2. By contrast, the use of VPs without ADPs enabled the estimation of a unique rotation for less than 52% of the images. It should be noted that the number of unique axes was the same in the SL-MH, SL-PB, SP360, and HoliCity datasets because we used the same random distribution for generating fisheye images. Table S3 shows the number of VP/ADP labels in each dataset. The diagonal directions of ADPs led

Table S2. Comparison of the distribution of the number of unique axes after the removal of label ambiguity in Section 4.2 (main paper) (%)

Dataset[1]	Number of unique axes							
	0	1	2	3	4	5	6	7
Only VPs (5 points)								
Train	0.9	48.4	31.5	19.2	–	–	–	–
Test	0.8	47.8	31.5	19.9	–	–	–	–
VPs and ADPs (13 points)								
Train	0.0	1.3	13.5	25.7	24.8	18.8	10.9	5.1
Test	0.0	1.4	12.8	25.7	25.6	19.6	10.2	4.6

[1] SL-MH, SL-PB, SP360, and HoliCity all have the same distribution of the number of unique axes, as shown in this table

Table S3. Distribution of the number of labels after the removal of label ambiguity in Section 4.2 (main paper) (%)

Label name[1]	Train[2]	Test[2]
VPs		
front	57.2	46.9
back	0.1	0.3
left	37.6	42.3
right	21.3	19.2
top	26.6	31.0
bottom	26.5	31.3
ADPs		
FLT	50.6	47.8
FRT	41.9	35.5
FLB	50.5	47.9
FRB	41.7	35.2
BLT	16.0	21.6
BRT	7.2	9.2
BLB	15.9	22.0
BRB	7.2	9.1

[1] The labels of the VPs and ADPs correspond to the labels described in Table 2 (main paper)
[2] SL-MH, SL-PB, SP360, and HoliCity all have the same distribution of the number of unique axes, as shown in this table

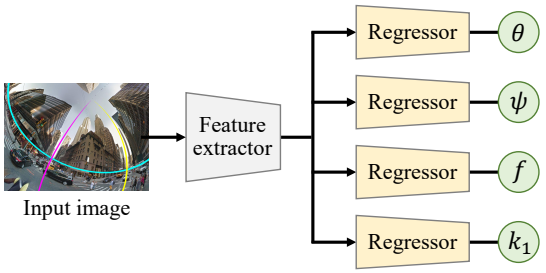
to increasing the number of ADPs in images. In particular, ADPs were arranged at the front side (FLT, FRT, FLB, and FRB) in over 35% of the images. Therefore, ADPs can compensate for the lack of VPs in images.

S5. Experimental results

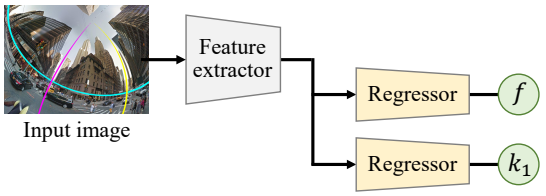
To demonstrate the validity and effectiveness of our method, we present further quantitative and qualitative results of our experiments in this section. The dataset names (SL-MH, SL-PB, SP360, and HoliCity) correspond to the names used in Section 4.1 (main paper).

S5.1. Results of training the distortion estimator

We report the performance of our distortion estimator to describe the difference between Wakai *et al.*'s method [18] and the distortion estimator. Our distortion estimator is composed of Wakai *et al.*'s calibration network [18] without the tilt and roll angle regressors, as shown in Figure S4. We optimized the distortion estimator after pretraining on Wakai *et al.*'s calibration network [18]. The distortion estimator achieved slight improvements in the focal length f and distortion coefficient k_1 because the number of estimated camera parameters was reduced by two, that is, the tilt and roll angles, as shown in Table S4.



(a) Wakai *et al.*'s calibration network



(b) Our distortion network

Figure S4. Network architectures of (a) Wakai *et al.*'s calibration network [18] and (b) our distortion network. Wakai *et al.*'s calibration network estimates extrinsics (tilt θ and roll ψ angles) and intrinsics (focal length f and a distortion coefficient k_1). By contrast, our distortion estimator has two regressors for the focal length f and distortion coefficient k_1. The input fisheye image is generated from [9].

S5.2. Comparison using ResNet backbones

To clarify the performance of the HRNet [16] backbones, we also evaluated our method using the ResNet [14] backbones, which are one of the baseline backbones used for various tasks. Table S5 shows the results of our method using either the ResNet or HRNet backbones. With respect to rotation errors and reprojection errors (REPE) [18], our method using the HRNet backbones outperformed that using the ResNet backbones, irrespective of backbone size. This benefit of the HRNet backbones corresponds to its advantages for human pose estimation [16]. Our method using HRNet-W48 achieved slight improvements over HRNet-W32 with respect to REPE and pan, tilt, and roll angles. The small magnitude of these improvements suggests that the performance is saturated for the larger HRNet-W48 backbone; this saturation may possibly be caused by the limitation of the variations of the panoramic-image datasets.

Table S4. Comparison of calibration accuracy by Wakai *et al.*'s method [18] and our distortion estimator on the test sets of each dataset

Dataset	Wakai *et al.* [18] ECCV'22						Our distortion estimator					
	Mean absolute error ↓				RSNR[1] ↑	SSIM[1] ↑	Mean absolute error ↓				RSNR ↑	SSIM ↑
	Tilt θ [deg]	Roll ψ [deg]	f [mm]	k_1			Tilt θ [deg]	Roll ψ [deg]	f [mm]	k_1		
SL-MH	4.13	5.21	0.34	0.021	29.01	0.838	–	–	0.34	**0.020**	**29.09**	**0.840**
SL-PB	4.06	5.71	0.36	0.024	29.05	0.826	–	–	0.36	**0.022**	**29.31**	**0.833**
SP360	3.75	5.19	0.39	0.023	28.10	0.835	–	–	**0.37**	0.023	**28.23**	**0.836**
HoliCity	6.55	16.05	0.48	0.028	25.59	0.751	–	–	0.48	0.028	**25.68**	**0.755**

[1] PSNR is the peak signal-to-noise ratio and SSIM is the structural similarity [20]

Table S5. Comparison of ResNet and HRNet in our method on the SL-MH test set

Backbone[1]	Mean absolute error[2] ↓			REPE[2] ↓	Mean fps[3] ↑	#Params	GFLOPs[4]
	Pan ϕ	Tilt θ	Roll ψ				
ResNet-50	4.89	4.97	4.79	8.39	**19.9**	58.7M	16.4
ResNet-101	3.65	4.07	3.87	7.02	17.7	76.8M	19.8
ResNet-152	3.46	3.80	3.72	6.68	16.0	91.8M	23.3
HRNet-W32	2.20	3.15	3.00	5.50	12.3	53.5M	14.5
HRNet-W48	**2.19**	**3.10**	**2.88**	**5.34**	12.2	86.9M	22.1

[1] Our VP estimator backbones are indicated
[2] Units: pan ϕ, tilt θ, and roll ψ [deg]; REPE [pixel]
[3] Implementations: our method using PyTorch [11]
[4] Rotation estimation in Figure 4 (main paper) is not included

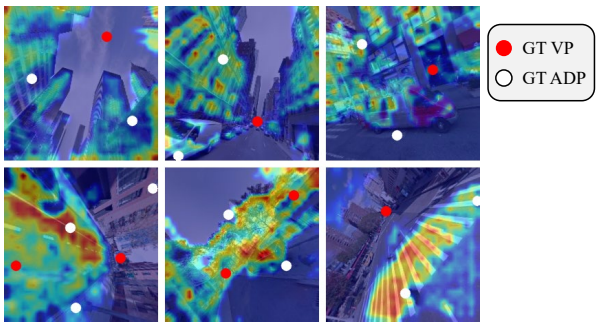


Figure S5. Eigen-CAM [2] results of our VP estimator (ResNet-50 [14]) on the SL-MH test set. Ground-truth (GT) VPs and ADPs are shown using red and white circles, respectively.

S5.3. Visualization of our VP estimator

To analyze the network activation of our VP estimator, we visualized the activation of the middle layers using Eigen-CAM [2]. For the visualization, ResNet-50 [14] backbones were used for simplicity because HRNet [16] backbones have branched structures. The ResNet-50 backbones without the head layer consist of 49 convolutional layers. These layers can be divided by sequential five blocks from input to output: conv1, conv2_x, conv3_x, conv4_x, and conv5_x. It should be noted that the features of conv5_x are used for the head of the deconvolutional layer block. Therefore, we selected conv2_x at the middle of the layers to analyze the network responses because the features of conv5_x are heatmaps with activated VP/ADPs.

Figure S5 shows the visualization of conv2_x of ResNet-50 backbones in our VP estimator using Eigen-CAM on the SL-MH test set. This result suggests that the VP estimator tends to extract image features from continuous textured regions, such as buildings, vehicles, and roads. The deformation of these continuous regions can have implicit 3D information used for the final deconvolution layer block to detect VP/ADPs.

S5.4. Comparison using HRNet loss function

To validate the effectiveness of our loss function, we trained the VP estimator with HRNet-W32 using the HRNet loss function [16]. As described in Section 3.2 (main paper), the HRNet loss function evaluates only images that include detected keypoints; that is, detection failure does not affect the loss value based on pixel values. To solve this problem, we modified this loss function to evaluate all images, including those with detection failure. Table S6 shows the results of our VP estimator trained using either the HRNet loss function or our loss function. In keypoint metrics, the VP estimator trained using our loss function improved the average precision (AP), average recall (AR)^{50}, and AR^{75} by 0.01 points, compared with the VP estimator trained using the HRNet loss function. The mean distance errors of all VP/ADPs in the VP estimator trained using our loss function are smaller than those in the VP estimator trained using the HRNet loss function by 0.08 pixels. In addition, the results in Table S7 show that our method trained using our loss function improved angle estimation by 0.17° on average, compared with our method using the HRNet loss function,

Table S6. Comparison of loss functions in our VP estimator using HRNet-W32 on the SL-MH test set

Loss function	Keypoint metric ↑							Mean distance error [pixel] ↓							
	AP	AP^{50}	AP^{75}	AR	AR^{50}	AR^{75}	PCK	front	left	right	top	bottom	VP^1	ADP^1	All^1
HRNet loss function [16]	0.98	0.99	0.99	0.97	0.97	0.97	0.99	**2.58**	**2.86**	2.55	1.90	**1.69**	**2.35**	3.80	3.18
Our loss function	**0.99**	0.99	0.99	0.97	**0.98**	**0.98**	0.99	2.67	2.90	**2.52**	1.90	1.72	2.39	**3.64**	**3.10**

[1] VP denotes all 5 VPs; ADP denotes all 8 ADPs; All denotes all points consisting of 5 VPs and 8 ADPs

Table S7. Comparison of loss functions in our method using HRNet-W32 on the SL-MH test set

Loss function	Mean absolute error[1] ↓			REPE[1] ↓
	Pan ϕ	Tilt θ	Roll ψ	
HRNet loss function [16]	2.54	3.25	3.07	5.60
Our loss function	**2.20**	**3.15**	**3.00**	**5.50**

[1] Units: pan ϕ, tilt θ, and roll ψ [deg]; REPE [pixel]

for pan, tilt, and roll angles. Therefore, our loss function can improve the calibration accuracy of our method.

S5.5. Error factor of our method

We analyzed the results of our method using HRNet-W32 to describe the error factor; that is, the calibration errors were caused by the distortion estimator and VP estimator. To evaluate this error factor, we performed calibration with ground-truth values for distortion parameters and image coordinates of VP/ADPs from the distortion estimator and the VP estimator, respectively, as shown in Table S8. Our method using ground-truth image coordinates of VP/ADPs outperformed our method using ground-truth distortion parameters, with respect to angle error and REPE. Therefore, the errors of the VP estimator were dominant over those of the distortion errors. In particular, angle errors were primarily caused by the VP estimator because camera rotation is mainly estimated from VP/ADPs. These results also show that the distortion estimator and VP estimator have room for improvement by $-0.11°$ and $-1.58°$, respectively, on average for pan, tilt, and roll angles.

S5.6. Details of quantitative results

To analyze the accuracy and robustness of our method, we evaluated our method and conventional methods on the test sets of SL-MH, SL-PB, SP360, and HoliCity. Table S9 shows the mean absolute errors and REPE. It should be noted that we cannot calculate the REPE of the Pritts *et al.*'s [12] and Lochman *et al.*'s [6] methods, for the following reasons: Pritts *et al.*'s method does not estimate focal length, which we need for calculating REPE; it is hard for Lochman *et al.*'s method using the division camera model [4] to address projected sampling points with over $180°$ fields of view because camera parameter errors lead to projected points with over $180°$ fields of view. Table S10 also reports the results of the cross-domain evaluation. These results demonstrated that our method using HRNet-W32 outperformed methods proposed by López-Antequera *et al.* [7], Wakai and Yamashita [17], Wakai *et al.* [18], Pritts *et al.* [12], and Lochman *et al.* [6] in terms of the mean absolute errors and REPE.

S5.7. Error distribution of our method

To evaluate the error distribution of angles, we compared the estimated and ground-truth camera parameters. Figure S6 shows the error distribution for our method using HRNet-W32. Although a few estimated angles have angle errors, most estimated angles are plotted close to the diagonal lines in Figure S6. (Angles are plotted on the diagonal lines when the estimated angles correspond to the ground-truth angles.) This distribution indicates that our method can stably estimate angles throughout the angle range from $-90°$ to $90°$; that is, it demonstrates angle robustness.

In addition, we analyzed the error distribution of camera parameters: angles, focal length, and distortion coefficients. We divided the angle range into 10 equal intervals: $[-90°, -72°]$, $[-72°, -54°]$, $\ldots$, $[72°, 90°]$. Similarly, we divided the ranges of focal length and distortion coefficients into 10 equal intervals. The results for these subdivisions are shown using box and violin[1] plots in Figure S7. Each violin plot with a single peak indicates that our networks were sufficiently optimized because insufficient optimization leads to multiple peaks in violin plots. Overall, our method achieved precise calibration across the whole range of estimated angles.

S5.8. Qualitative evaluation

To validate the VP/ADP estimation and quality of the recovered images, we present additional calibration results using synthetic images and off-the-shelf fisheye cameras.

[1] The violin plot represents the probability density of the distribution as the width of the violin plot, and supports multiple peaks. Two peaks of the probability density form the shape of a violin.

Table S8. Comparison using estimation and ground truth in our method using HRNet-W32 on the SL-MH test set

Distortion parameter f and k_1 (Distortion estimator)	Image coordinates of VP/ADPs (VP estimator)	Mean absolute error[1] ↓			REPE[1] (Gain) ↓
		Pan ϕ (Gain[2])	Tilt θ (Gain)	Roll ψ (Gain)	
Estimation	GT	**0.94** (−1.26)	**1.40** (−1.75)	**1.27** (−1.73)	**2.60** (−2.90)
GT	Estimation	2.21 (+0.01)	2.95 (−0.20)	2.86 (−0.14)	3.83 (−1.67)
Estimation	Estimation	2.20	3.15	3.00	5.50

[1] Units: pan ϕ, tilt θ, and roll ψ [deg]; REPE [pixel]
[2] The origin of the gain is that our method estimates both distortion parameters and image coordinates of VP/ADPs (bottom row)

Table S9. Comparison of the absolute parameter errors and reprojection errors on the test sets of each dataset

Dataset	Method		Mean absolute error[1] ↓					REPE[1] ↓	Executable rate[1] ↑
			Pan ϕ	Tilt θ	Roll ψ	f	k_1		
SL-MH	López-Antequera *et al.* [7]	CVPR'19	–	27.60	44.90	2.32	–	81.99	100.0
	Wakai and Yamashita [17]	ICCVW'21	–	10.70	14.97	2.73	–	30.02	100.0
	Wakai *et al.* [18]	ECCV'22	–	4.13	5.21	0.34	0.021	7.39	100.0
	Pritts *et al.* [12]	CVPR'18	25.35	42.52	18.54	–	–	–	96.7
	Lochman *et al.* [6]	WACV'21	22.36	44.42	33.20	6.09	–	–	59.1
	Ours (HRNet-W32)		**2.20**	**3.15**	**3.00**	0.34	0.020	**5.50**	100.0
SL-PB	López-Antequera *et al.* [7]	CVPR19	–	26.18	41.94	2.11	–	73.68	100.0
	Wakai and Yamashita [17]	ICCVW'21	–	10.66	14.53	2.67	–	25.76	100.0
	Wakai *et al.* [18]	ECCV'22	–	4.06	5.71	0.36	0.024	7.99	100.0
	Pritts *et al.* [12]	CVPR'18	25.55	42.94	18.28	–	–	–	97.9
	Lochman *et al.* [6]	WACV'21	23.45	44.99	30.68	8.14	–	–	39.1
	Ours (HRNet-W32)		**2.30**	**3.13**	**3.09**	0.36	0.022	**5.89**	100.0
SP360	López-Antequera *et al.* [7]	CVPR19	–	28.66	44.45	3.26	–	84.56	100.0
	Wakai and Yamashita [17]	ICCVW'21	–	11.12	17.70	2.67	–	32.01	100.0
	Wakai *et al.* [18]	ECCV'22	–	3.75	5.19	0.39	0.023	7.39	100.0
	Pritts *et al.* [12]	CVPR'18	25.39	42.79	18.35	–	–	–	98.5
	Lochman *et al.* [6]	WACV'21	22.84	45.38	31.91	6.81	–	–	53.7
	Ours (HRNet-W32)		**2.16**	**2.92**	**2.79**	0.37	0.023	**5.60**	100.0
HoliCity	López-Antequera *et al.* [7]	CVPR'19	–	65.92	50.31	2.27	–	96.63	100.0
	Wakai and Yamashita [17]	ICCVW'21	–	12.18	26.00	2.56	–	34.99	100.0
	Wakai *et al.* [18]	ECCV'22	–	6.55	16.05	0.48	0.028	19.37	100.0
	Pritts *et al.* [12]	CVPR'18	25.45	43.22	17.84	–	–	–	99.6
	Lochman *et al.* [6]	WACV'21	22.63	45.11	32.58	6.71	–	–	83.9
	Ours (HRNet-W32)		**3.48**	**4.08**	**3.84**	0.48	0.028	**7.62**	100.0

[1] Units: pan ϕ, tilt θ, and roll ψ [deg]; f [mm]; k_1 [dimensionless]; REPE [pixel]; Executable rate [%]

S5.8.1 Vanishing point estimation

As described in Section 4.4.1 (main paper), the VP estimator detected the VP/ADPs, although the performance in the cross-domain evaluation decreased in Table S11. In addition, Table S11 reveals that ADP detection is more difficult than VP detection because VPs generally have specific appearances at infinity. In the cross-domain evaluation, models trained by HoliCity could adapt well to other domains.

To demonstrate the robustness of our heatmap-based VP estimator, we visualized the results of VP/ADPs. Figure S8 shows qualitative results of the VP estimator using HRNet-W32. Although the test images were affected by various types of rotation and distortion, the VP estimator achieved stable VP/ADP detection from the centers of images to their edges. Each estimated VP/ADP heatmap has a single peak for VP/ADPs. Such a single peak, with little noise, indicates that the VP estimator was well-optimized. In addition, many test images contain large regions of sky or road surface with few geometric cues such as arcs; however, the VP estimator handled these images successfully. Therefore, the VP estimator was able to robustly detect VP/ADPs.

S5.8.2 Recovered images

Synthetic images. Figure S9 shows the additional qualitative results obtained on synthetic images. Similarly to Figure 6 (main paper), our results are the most similar to the ground-truth images. By contrast, the quality of the recovered images that contain a few arcs was notably degraded when the geometry-based methods proposed by Pritts *et al.* [12] and Lochman *et al.* [6] were used. In particular, Lochman *et al.*'s method [6] tended to result in execution failure on these images. Additionally, the learning-based methods proposed by López-Antequera *et al.* [7], Wakai and Yamashita [17], and Wakai *et al.* [18] did not recover the pan angles; that is, vertical magenta and yellow lines are not located at the centers of the images produced by these methods, shown in Figure S9. We note that our method was able to calibrate images of streets lined by large trees.

Table S10. Comparison on the cross-domain evaluation of the mean absolute rotation errors and reprojection errors

Dataset		López-Antequera *et al*. [7] CVPR'19				Wakai and Yamashita [17] ICCVW'21				Wakai *et al*. [18] ECCV'22				Ours (HRNet-W32)			
Train	Test	Pan[1]	Tilt[1]	Roll[1]	REPE[1]	Pan	Tilt	Roll	REPE	Pan	Tilt	Roll	REPE	Pan	Tilt	Roll	REPE
SL-MH	SL-PB	–	31.11	45.16	83.42	–	12.99	27.13	39.43	–	5.51	12.02	14.89	2.98	**3.72**	**3.63**	**6.82**
	SP360	–	28.91	45.23	82.68	–	12.29	38.42	55.72	–	9.11	37.54	43.56	8.06	**8.34**	**7.77**	**17.85**
	HoliCity	–	33.36	45.20	82.40	–	13.78	45.76	53.99	–	10.94	42.20	47.97	10.74	**10.60**	**8.93**	**19.84**
SL-PB	SL-MH	–	26.92	46.35	76.09	–	11.65	26.50	36.48	–	5.18	13.77	16.99	3.04	**3.58**	**3.39**	**6.68**
	SP360	–	28.29	48.10	78.87	–	12.57	40.25	47.50	–	9.61	40.05	46.07	8.78	**8.93**	**8.28**	**18.66**
	HoliCity	–	32.64	50.37	80.98	–	13.79	46.06	51.72	–	12.53	42.77	49.43	10.95	**11.17**	**9.47**	**20.62**
SP360	SL-MH	–	32.44	47.18	90.97	–	16.25	41.12	49.02	–	8.72	38.96	47.89	6.52	**6.82**	**6.52**	**15.66**
	SL-PB	–	34.31	46.63	90.99	–	16.07	38.38	51.72	–	7.42	37.09	45.45	5.18	**5.81**	**5.60**	**14.11**
	HoliCity	–	30.84	49.19	83.43	–	16.66	44.42	55.47	–	12.83	43.81	51.26	12.65	**12.11**	**10.41**	**20.48**
HoliCity	SL-MH	–	65.52	50.41	96.29	–	14.20	35.44	46.32	–	8.97	33.35	40.48	6.13	**6.54**	**5.97**	**14.72**
	SL-PB	–	65.69	50.95	96.84	–	15.00	47.07	56.54	–	9.59	42.28	49.38	5.26	**5.88**	**5.73**	**14.85**
	SP360	–	64.43	51.59	96.59	–	13.67	42.39	50.36	–	9.43	37.83	43.59	6.10	**6.57**	**6.37**	**13.31**

[1] Units: pan, tilt, and roll [deg]; REPE [pixel]

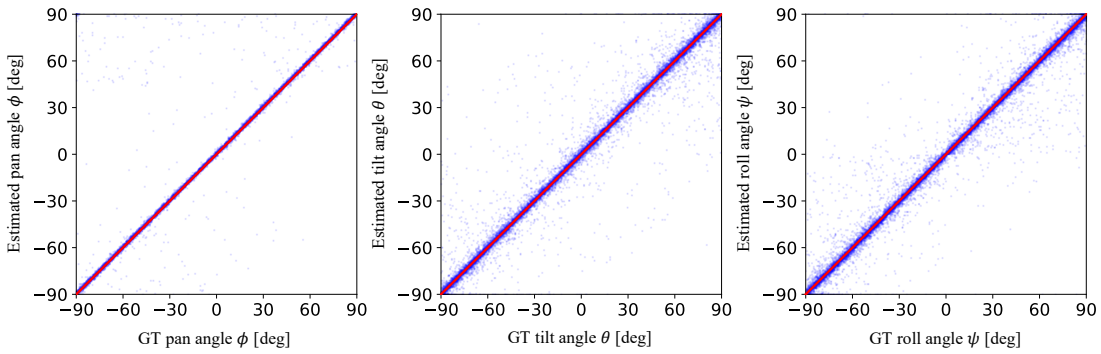
Table S11. Results of the cross-domain evaluation for our VP estimator using HRNet-W32

Dataset		Keypoint metric ↑							Mean distance error [pixel] ↓							
Train	Test	AP	AP^{50}	AP^{75}	AR	AR^{50}	AR^{75}	PCK	front	left	right	top	bottom	VP[1]	ADP[1]	All[1]
SL-MH	SL-MH	0.99	0.99	0.99	0.97	0.98	0.98	0.99	2.67	2.90	2.52	1.90	1.72	2.39	3.64	3.10
	SL-PB	0.98	0.99	0.99	0.96	0.97	0.97	0.98	3.51	3.50	3.11	2.34	2.02	2.97	4.52	3.85
	SP360	0.85	0.94	0.90	0.79	0.87	0.83	0.83	6.55	7.42	6.18	5.34	11.77	7.44	14.95	11.57
	HoliCity	0.80	0.92	0.86	0.72	0.83	0.78	0.77	9.73	12.27	9.75	8.54	6.60	9.47	17.92	14.11
SL-PB	SL-MH	0.99	0.99	0.99	0.96	0.97	0.97	0.99	3.26	3.49	3.10	2.04	1.74	2.79	4.63	3.84
	SL-PB	0.99	0.99	0.99	0.97	0.97	0.97	0.99	2.91	2.93	2.48	1.97	1.80	2.49	3.68	3.17
	SP360	0.82	0.92	0.87	0.75	0.85	0.81	0.81	7.72	8.65	7.53	5.41	12.74	8.42	15.88	12.53
	HoliCity	0.77	0.91	0.83	0.70	0.82	0.76	0.74	11.33	13.33	11.34	10.63	7.14	10.84	19.49	15.60
SP360	SL-MH	0.95	0.98	0.97	0.89	0.91	0.91	0.94	5.13	5.38	4.46	4.24	4.67	4.88	9.78	7.63
	SL-PB	0.95	0.98	0.97	0.88	0.91	0.90	0.94	4.66	4.86	4.12	4.83	3.69	4.50	10.14	7.66
	SP360	0.99	1.00	1.00	0.98	0.98	0.98	0.99	2.64	2.61	2.37	1.78	1.80	2.29	3.20	2.81
	HoliCity	0.79	0.92	0.85	0.69	0.79	0.75	0.77	12.44	14.20	11.33	8.20	6.02	10.70	19.38	15.45
HoliCity	SL-MH	0.95	0.98	0.97	0.89	0.91	0.91	0.95	4.80	5.07	4.44	3.89	4.50	4.61	10.50	7.89
	SL-PB	0.95	0.98	0.97	0.89	0.91	0.91	0.95	5.11	5.00	4.22	4.67	3.65	4.64	10.23	7.76
	SP360	0.89	0.96	0.93	0.84	0.90	0.87	0.88	5.42	6.75	6.05	3.34	9.59	6.20	12.11	9.48
	HoliCity	0.98	0.99	0.99	0.95	0.96	0.96	0.98	3.44	3.87	3.30	3.30	2.66	3.36	5.70	4.68

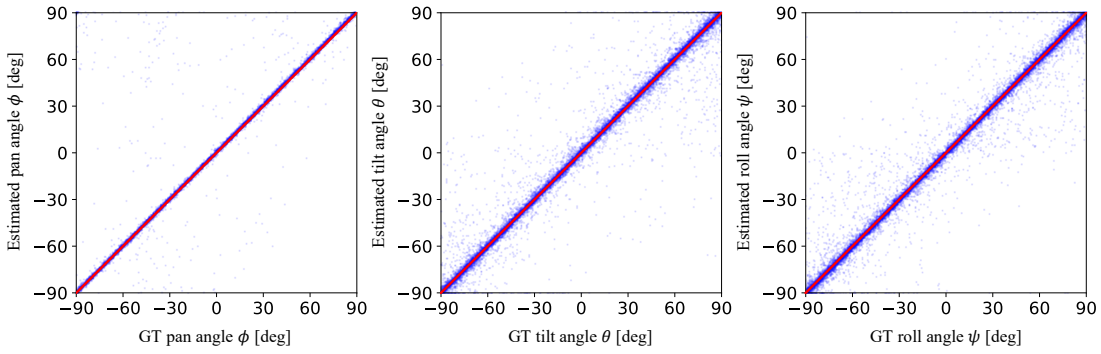
[1] VP denotes all 5 VPs; ADP denotes all 8 ADPs; All denotes all points consisting of 5 VPs and 8 ADPs

To validate the effectiveness of our method, we also demonstrated the qualitative results in the cross-domain evaluation. Figure S10 shows the qualitative results in the cross-domain evaluation on the HoliCity test set when learning-based methods were trained on SL-MH. Conventional learning-based methods tended to have rotation errors in the cross-domain evaluation, as shown in Table S10. We found that Wakai *et al*.'s method [18] often recovered images upside down in a cloudy sky. This observation suggests that regression-based methods that do not use heatmaps, such as Wakai *et al*.'s method [18], tend to misinterpret the cloudy sky as a gray road. It should be noted that the images in HoliCity were captured in London, where the weather is often cloudy all year round. This phenomenon implies that regression-based methods that do not use heatmaps estimate the roll angles mainly based on the sky and road regions. Although the sky and roads generally occupy large areas of these regions, which have fewer geometric cues, seem to lead to unstable estimation. By contrast, our method, which uses heatmaps, can extract robust features through geometric VP/ADPs. Therefore, our method achieved robust estimation in various domains.

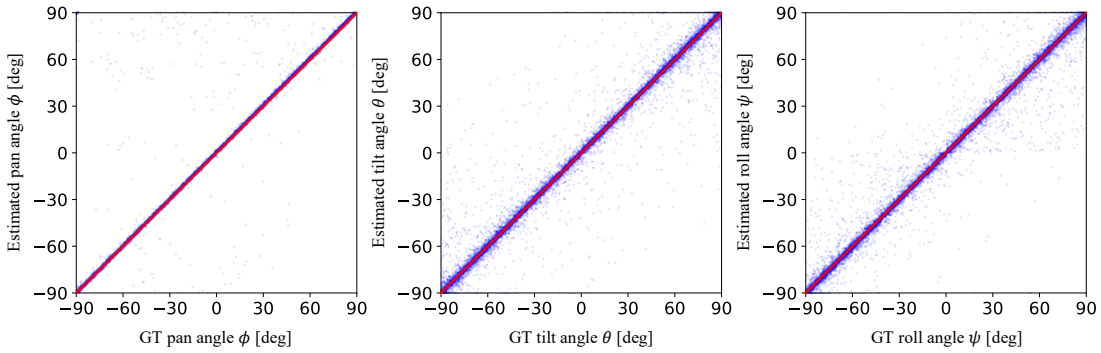
Off-the-shelf cameras. Following [18], we also evaluated calibration methods using six off-the-shelf fisheye cameras to validate the effectiveness of our method. Figure S11 shows the qualitative results on images from off-the-shelf fisheye cameras using SL-MH for training. Similarly to Figure 7 (main paper), our method substantially outperformed the methods proposed by López-Antequera *et al*. [7], Wakai and Yamashita [17], Wakai *et al*. [18], Pritts *et al*. [12], and Lochman *et al*. [6] with respect to the quality of the recovered images. Furthermore, these results demonstrate the robustness of our method for four types of camera projection: equisolid angle projection, orthogonal projection, equidistant projection, and stereographic projection. A promising direction for future work is to quantitatively evaluate our method using off-the-shelf fisheye cameras in various scenes.



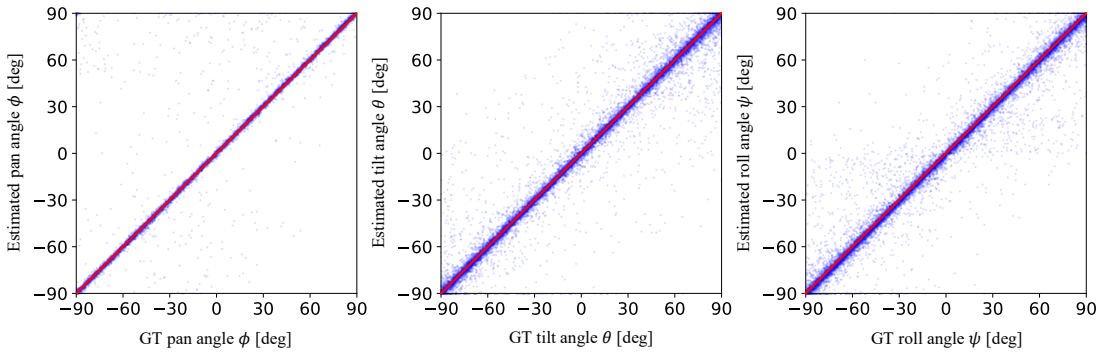
(a) SL-MH



(b) SL-PB

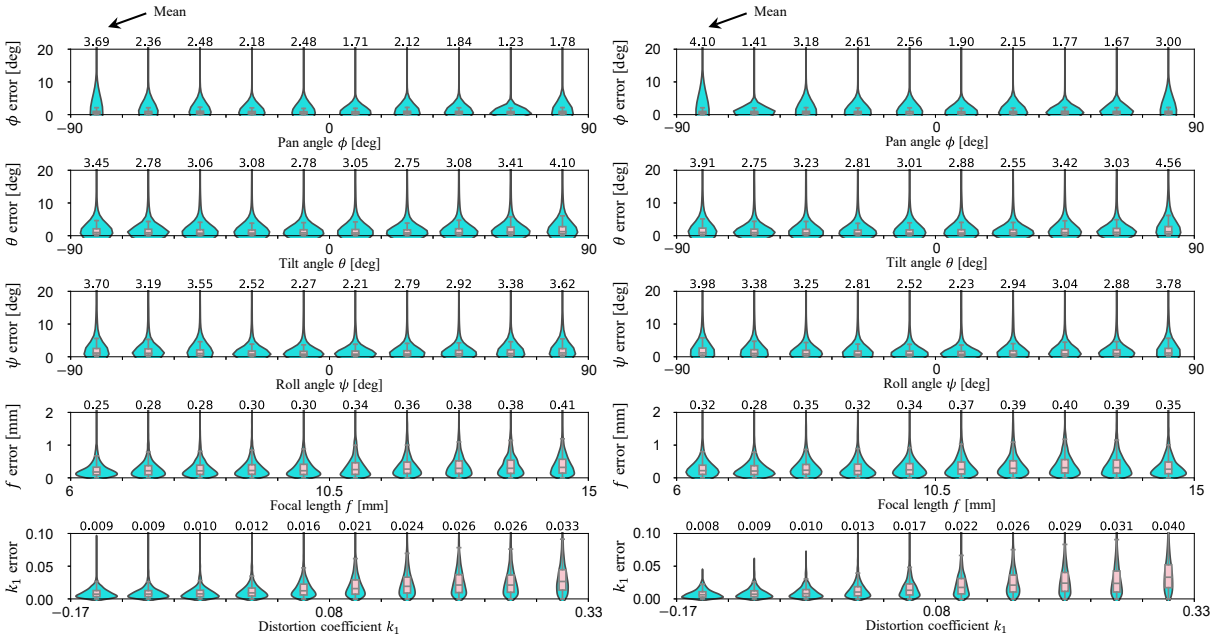


(c) SP360



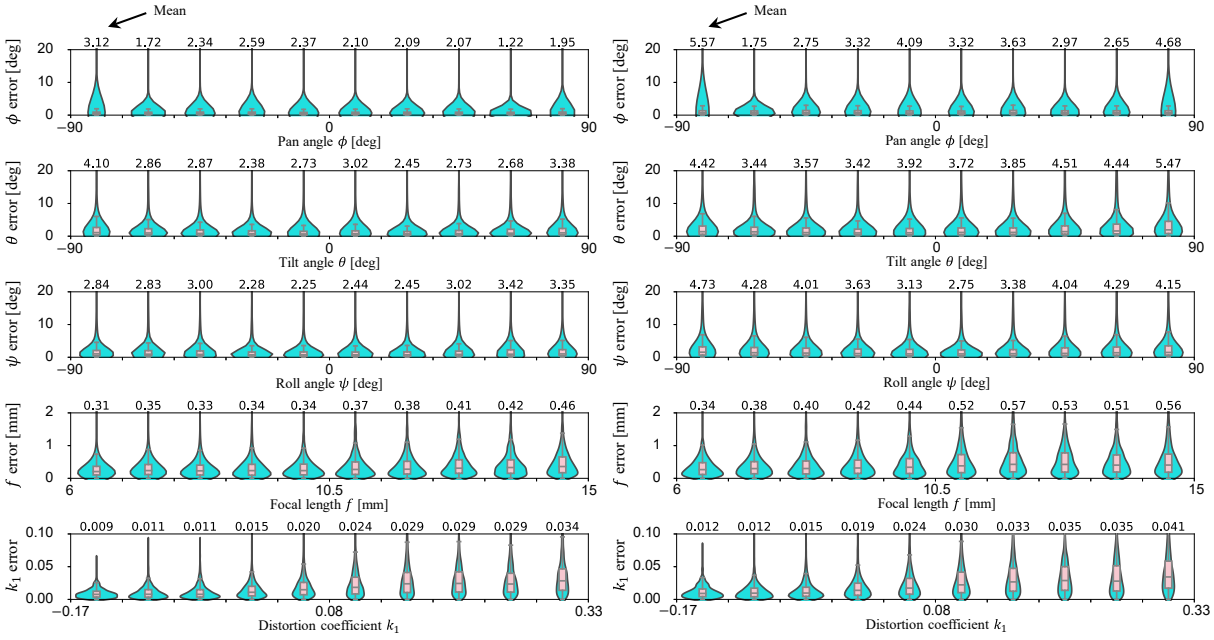
(d) HoliCity

Figure S6. Error distribution of angles in our method using HRNet-W32 on the test sets of (a) SL-MH, (b) SL-PB, (c) SP360, and (d) HoliCity. Ground-truth and estimated angles are indicated on the horizontal and vertical axes, respectively. The diagonal red lines represent perfect estimation without angle errors. Each estimation result for the test images is depicted as a translucent blue point.



(a) SL-MH

(b) SL-PB



(c) SP360

(d) HoliCity

Figure S7. Error distribution of our method on the test sets of (a) SL-MH, (b) SL-PB, (c) SP360, and (d) HoliCity. Mean absolute errors are shown in divided ranges as box and violin plots.

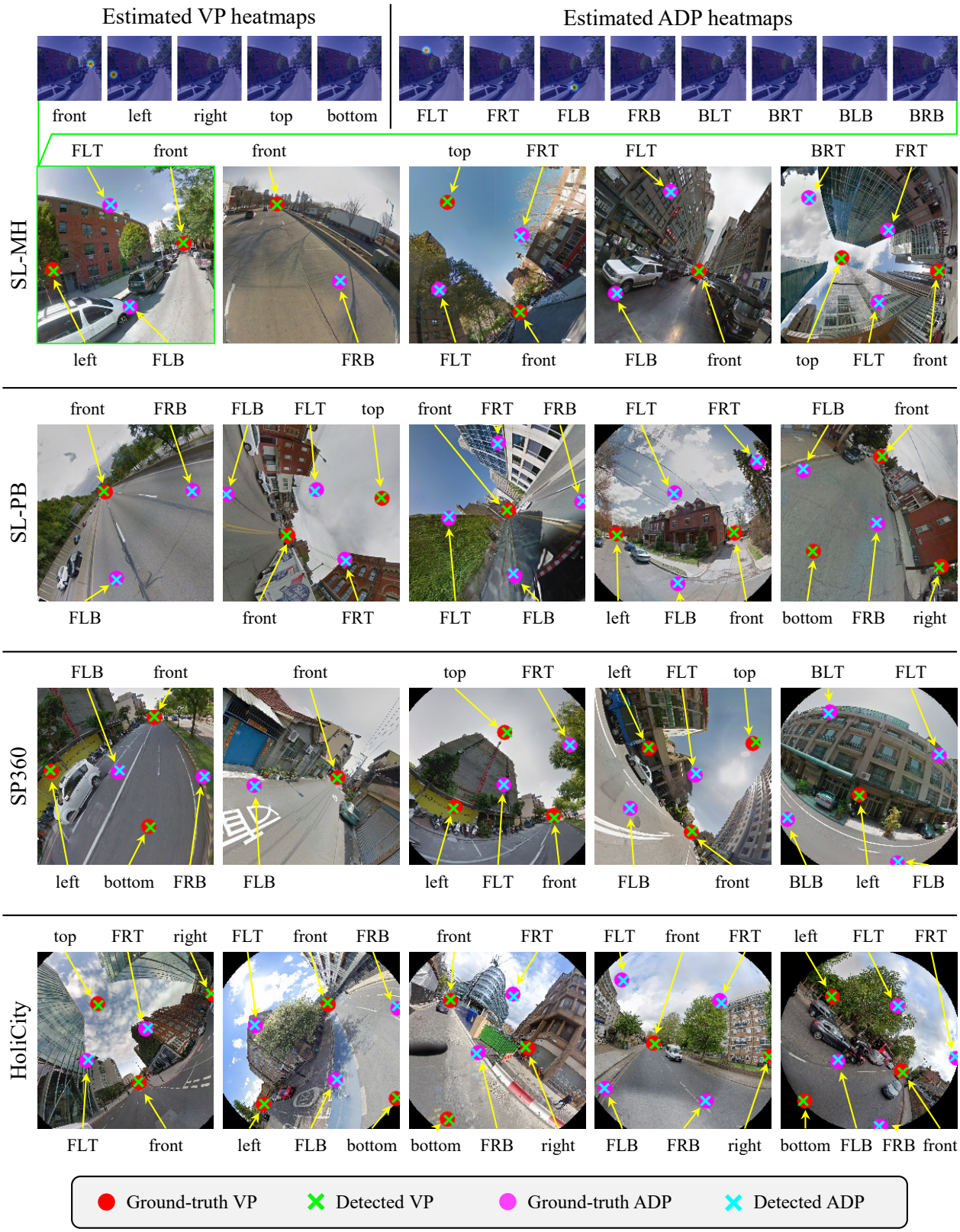
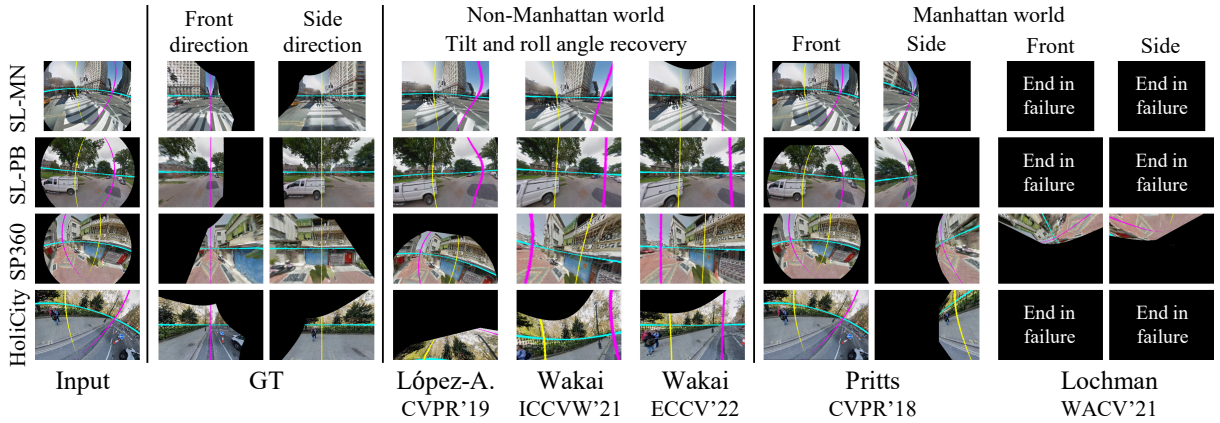
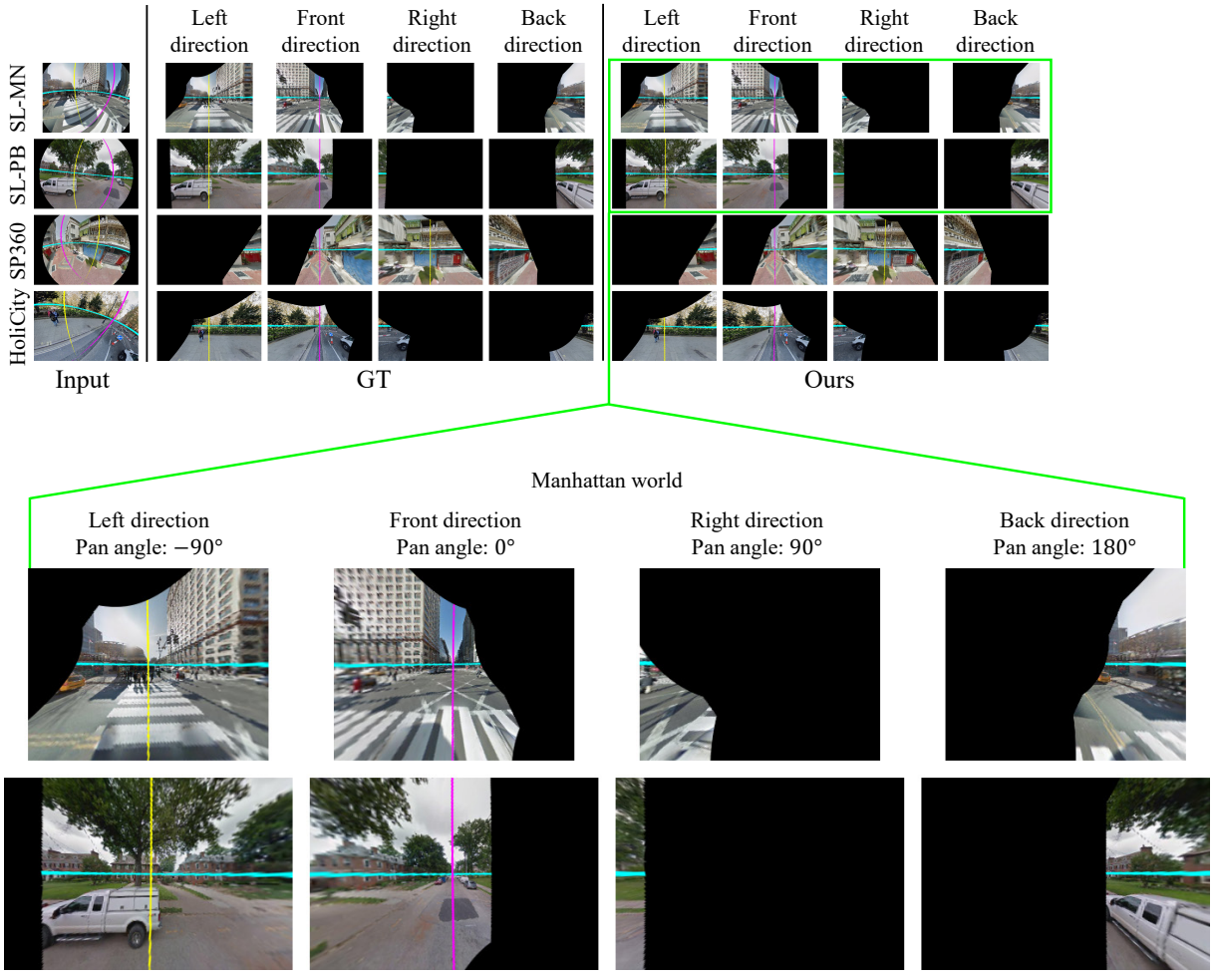


Figure S8. Qualitative results of VP/ADP estimation performed by our VP estimator using HRNet-W32 on the test sets of each dataset. The VP/ADP labels correspond to the labels in Table 2 (main paper).



(a) Results of conventional methods



(b) Results of our method

Figure S9. Qualitative results on the test sets. (a) Results of conventional methods. From left to right: input images, GT images, and results of López-Antequera *et al*. [7], Wakai and Yamashita [17], Wakai *et al*. [18], Pritts *et al*. [12], and Lochman *et al*. [6]. (b) Results of our method. From left to right: input images, GT images, and the results of our method using HRNet-W32 in a Manhattan world.

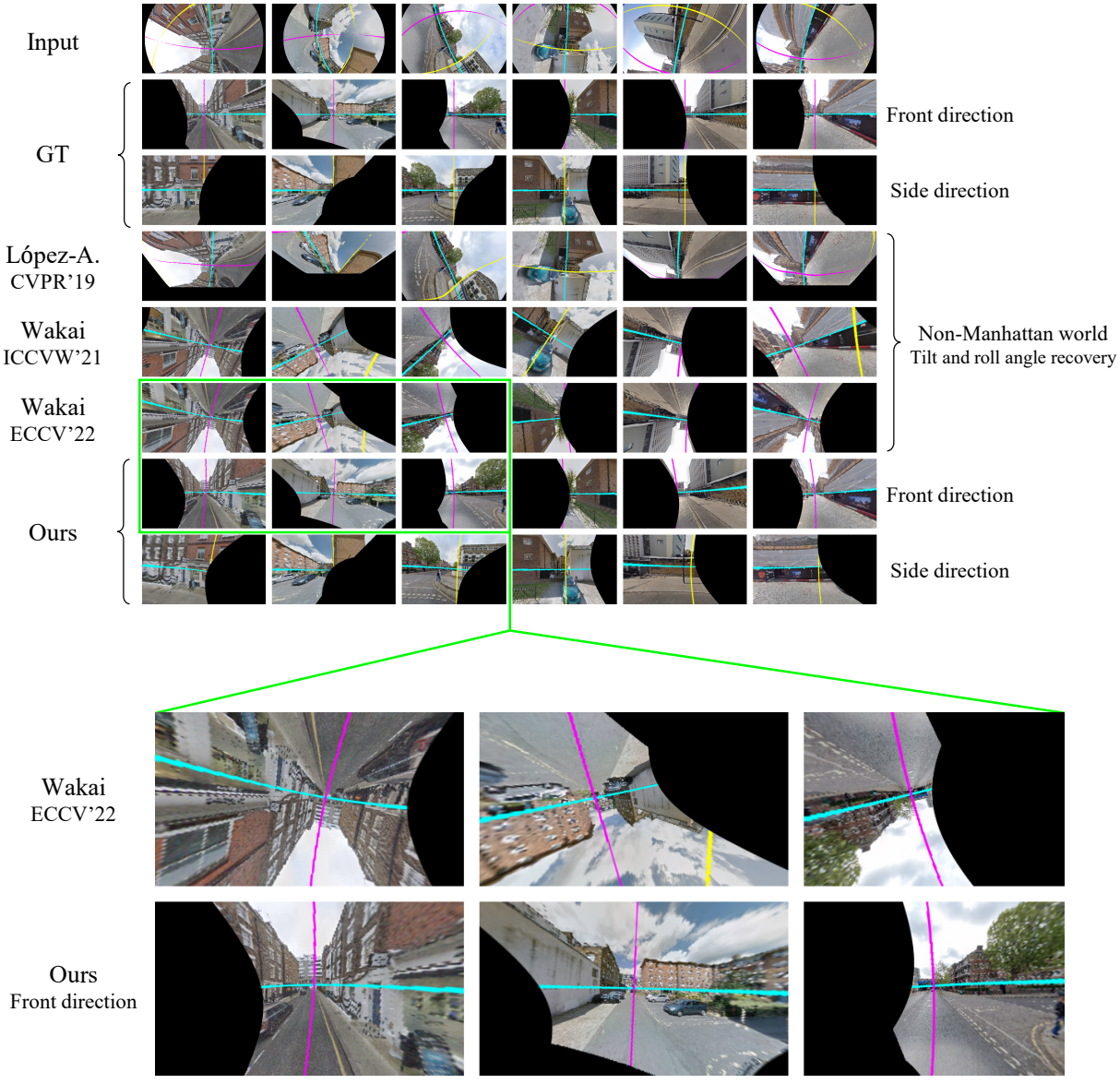


Figure S10. Qualitative results in the cross-domain evaluation on the HoliCity test set. Our method using HRNet-W32 and compared methods were trained on SL-MH. From top to bottom: input images, ground-truth images, and results of López-Antequera *et al*. [7], Wakai and Yamashita [17], Wakai *et al*. [18], and our method.

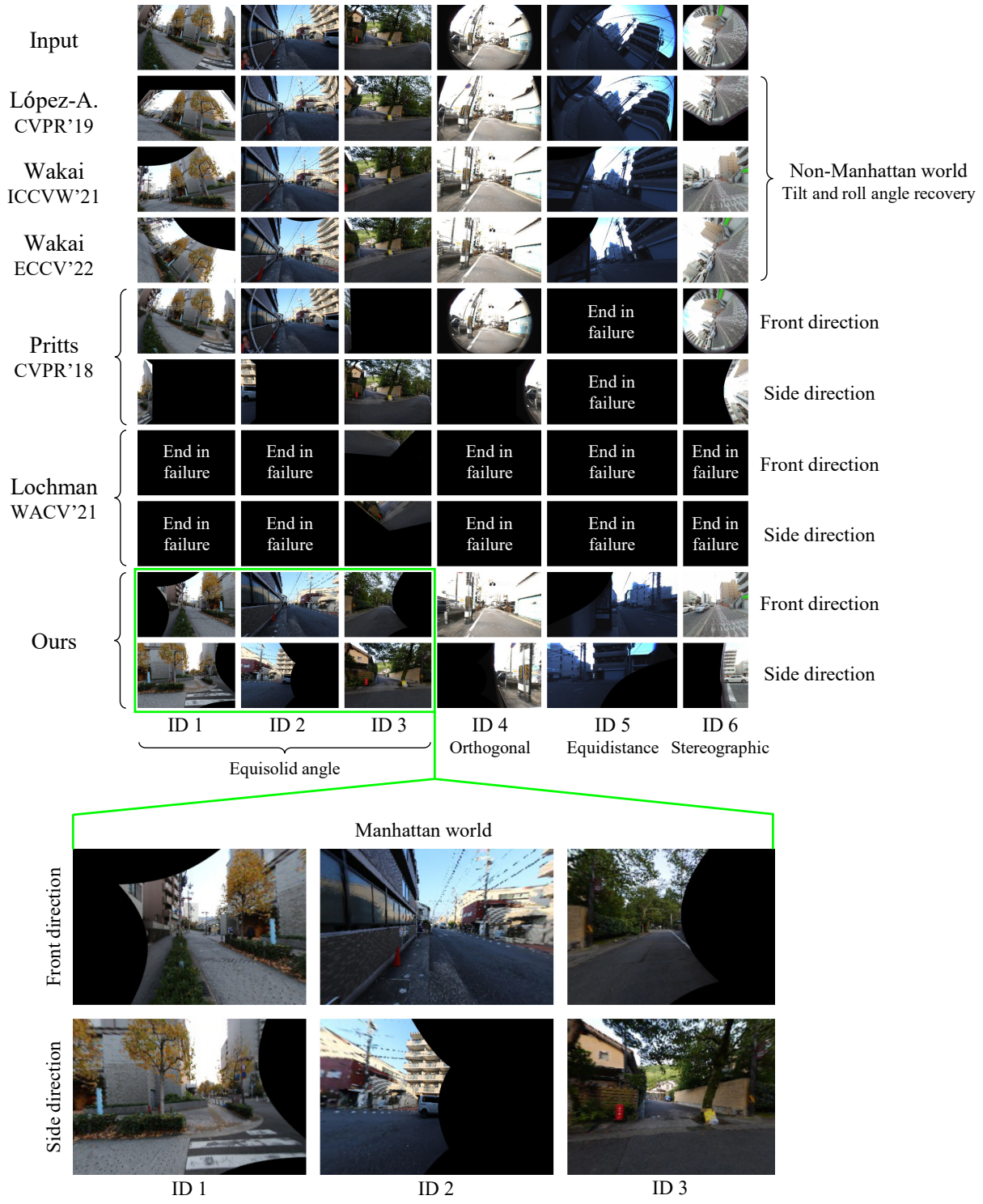


Figure S11. Qualitative results for images from off-the-shelf cameras. From top to bottom: input images and results of López-Antequera *et al*. [7], Wakai and Yamashita [17], Wakai *et al*. [18], Pritts *et al*. [12], Lochman *et al*. [6], and our method. The identifiers (IDs) correspond to the camera IDs used in [18], and the projection names are shown below the IDs.

References

[1] M. Antunes, J. P. Barreto, D. Aouada, and B. Ottersten. Unsupervised vanishing point detection and camera calibration from a single Manhattan image with radial distortion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6691–6699, 2017. 1

[2] M. Bany Muhammad and M. Yeasin. Eigen-CAM: Visual explanations for deep convolutional neural networks. *SN Computer Science*, 2(47), 2021. 6

[3] M. Eder, M. Shvets, J. Lim, and J. Frahm. Tangent images for mitigating spherical distortion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12423–12431, 2020. 2

[4] A. W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I125–I132, 2001. 7

[5] Y. Kang, Y. Song, W. Ge, and T. Ling. Robust multi-camera SLAM with Manhattan constraint toward automated valet parking. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7615–7622, 2021. 1

[6] Y. Lochman, O. Dobosevych, R. Hryniv, and J. Pritts. Minimal solvers for single-view lens-distorted camera auto-calibration. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2886–2895, 2021. 1, 7, 8, 9, 13, 15

[7] M. López-Antequera, R. Marí, P. Gargallo, Y. Kuang, J. Gonzalez-Jimenez, and G. Haro. Deep single image camera calibration with radial distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11809–11817, 2019. 1, 2, 7, 8, 9, 13, 14, 15

[8] M. Lourakis and G. Terzakis. Efficient absolute orientation revisited. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5813–5818, 2018. 3

[9] P. Mirowski, A. Banki-Horvath, K. Anderson, D. Teplyashin, K. M. Hermann, M. Malinowski, M. K. Grimes, K. Simonyan, K. Kavukcuoglu, A. Zisserman, and R. Hadsell. The StreetLearn environment and dataset. *arXiv preprint arXiv:1903.01292*, 2019. 5

[10] D. Mortari, F. L. Markley, and P. Singla. Optimal linear attitude estimator. *Journal of Guidance, Control, and Dynamics (JGCD)*, 3:1619–1627, 2007. 3

[11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019. 6

[12] J. Pritts, Z. Kukelova, V. Larsson, and O. Chum. Radially-distorted conjugate translations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1993–2001, 2018. 1, 7, 8, 9, 13, 15

[13] Z. Shen, C. Lin, K. Liao, L. Nie, Z. Zheng, and Y. Zhao. PanoFormer: Panorama transformer for indoor 360° depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 195–211, 2022. 2

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 5, 6

[15] C. Sun, M. Sun, and H. Chen. HoHoNet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2573–2582, 2021. 2

[16] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, 2019. 5, 6, 7

[17] N. Wakai and T. Yamashita. Deep single fisheye image camera calibration for over 180-degree projection of field of view. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1174–1183, 2021. 1, 2, 7, 8, 9, 13, 14, 15

[18] N. Wakai, S. Sato, Y. Ishii, and T. Yamashita. Rethinking generic camera models for deep single image camera calibration to recover rotation and fisheye distortion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 679–698, 2022. 1, 2, 5, 6, 7, 8, 9, 13, 14, 15

[19] F. Wang, Y. Yeh, M. Sun, W. Chiu, and Y. Tsai. BiFuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2020. 2

[20] Z. Wang and A. C. Bovik. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004. 6

[21] Z. Wang and Jepson. A new closed-form solution for absolute orientation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 129–134, 1994. 2

[22] H. Wildenauer and B. Micusik. Closed form solution for radial distortion estimation from a single vanishing point. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 106.1–106.11, 2013. 1

[23] Z. Yan, X. Li, L. Wang, Z. Zhang, J. Li, and J. Yang. Multi-modal masked pre-training for monocular panoramic depth completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 378–395, 2022. 2

[24] Z. Yan, X. Li, K. Wang, S. Chen, J. Li, and J. Yang. Distortion and uncertainty aware loss for panoramic depth completion. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 39099–39109, 2023. 2

[25] R. Yunus, Y. Li, and F. Tombari. ManhattanSLAM: Robust planar tracking and mapping leveraging mixture of Manhattan frames. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 6687–6693, 2021. 1