Table 6. 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

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

Table 7. Comparison of the absolute parameter errors and reprojection errors on the SL-MH test set

Method		Backbone	Mean absolute error[1] ↓					REPE[1] ↓	Executable rate[1] ↑	Mean fps[2] ↑	#Params	GFLOPs
			Pan ϕ	Tilt θ	Roll ψ	f	k_1					
López-Antequera *et al.* [33]	CVPR'19	DenseNet-161	–	27.60	44.90	2.32	–	81.99	100.0	36.4	27.4M	7.2
Wakai and Yamashita [52]	ICCVW'21	DenseNet-161	–	10.70	14.97	2.73	–	30.02	100.0	33.0	26.9M	7.2
Wakai *et al.* [53]	ECCV'22	DenseNet-161	–	4.13	5.21	0.34	0.021	7.39	100.0	25.4	27.4M	7.2
Pritts *et al.* [41]	CVPR'18	–	25.35	42.52	18.54	–	–	–	96.7	0.044	–	–
Lochman *et al.* [32]	WACV'21	–	22.36	44.42	33.20	6.09	–	–	59.1	0.016	–	–
Ours w/o ADPs	(5 points)[3]	HRNet-W32[3]	19.38	13.54	21.65	0.34	0.020	28.90	100.0	12.7	53.5M	14.5[3]
Ours w/o VPs	(8 points)	HRNet-W32	10.54	11.01	8.11	0.34	0.020	19.70	100.0	12.6	53.5M	14.5
Ours	(13 points)	HRNet-W32	**2.20**	**3.15**	**3.00**	0.34	0.020	**5.50**	100.0	12.3	53.5M	14.5
Ours	(13 points)	HRNet-W48	**2.19**	**3.10**	**2.88**	0.34	0.020	**5.34**	100.0	12.2	86.9M	22.1

[1] Units: pan ϕ, tilt θ, and roll ψ [deg]; f [mm]; k_1 [dimensionless]; REPE [pixel]; Executable rate [%]
[2] Implementations: López-Antequera [33], Wakai [52], Wakai [53], and ours using PyTorch [40]; Pritts [41] and Lochman [32] using The MathWorks MATLAB
[3] ($\cdot$ points) is the number of VP/ADPs for VP estimators; VP estimator backbones are indicated; Rotation estimation in Figure 4 is not included in GFLOPs

Table 8. Comparison of the mean absolute rotation errors in degrees on the test sets of each dataset

Dataset	Wakai *et al.* [53]			Lochman *et al.* [32]			Ours (HRNet-W32)		
	Pan	Tilt	Roll	Pan	Tilt	Roll	Pan	Tilt	Roll
SL-MH	–	4.13	5.21	22.36	44.42	33.20	**2.20**	**3.15**	**3.00**
SL-PB	–	4.06	5.71	23.45	44.99	30.68	**2.30**	**3.13**	**3.09**
SP360	–	3.75	5.19	22.84	45.38	31.91	**2.16**	**2.92**	**2.79**
HoliCity	–	6.55	16.05	22.63	45.11	32.58	**3.48**	**4.08**	**3.84**

Table 9. Comparison on the cross-domain evaluation of the mean absolute rotation errors in degrees

Dataset		Wakai *et al.* [53]			Ours (HRNet-W32)		
Train	Test	Pan	Tilt	Roll	Pan	Tilt	Roll
SL-MH	SL-PB	–	5.51	12.02	2.98	**3.72**	**3.63**
	SP360	–	9.11	37.54	8.06	**8.34**	**7.77**
	HoliCity	–	10.94	42.20	10.74	**10.60**	**8.93**

dress arbitrary images independent of the number of arcs; that is, it demonstrates scene robustness. Compared with methods [32, 41] estimating the pan angles, our method using HRNet-W32 achieved a mean frames per second (fps) that was at least 280 times higher. Note that our test platform was equipped with an Intel Core i7-6850K CPU and an NVIDIA GeForce RTX 3080Ti GPU.

We validated the effectiveness of the ADPs. Table 7 suggests that our method based on HRNet-W32 and VP/ADPs notably improved angle estimation compared with our method without the ADPs by $15.4°$ on average for pan, tilt, and roll angles. Therefore, the ADPs dramatically alleviated the problems caused by a lack of VPs.

Additionally, we tested our proposed method using various datasets to validate its robustness. Table 8 shows that our method outperforms both existing state-of-the-art learning-based [53] and geometry-based [32] methods on all datasets in terms of rotation errors. Table 9 also reports that our method is superior to Wakai *et al.*'s method [53], which tended to estimate the roll angle poorly in the cross-domain evaluation, especially on the HoliCity test set.

4.4.3 Qualitative evaluation

To evaluate the recovered image quality, we performed calibration on synthetic images and off-the-shelf cameras.

Synthetic images. Figure 6 shows the qualitative results obtained on synthetic images. Our results are the most similar to the ground-truth images. By contrast, the quality of the recovered images that contained a few arcs was considerably degraded when the geometry-based methods proposed by Pritts *et al.* [41] and Lochman *et al.* [32] were used. Furthermore, the learning-based methods proposed by López-Antequera *et al.* [33], Wakai and Yamashita [52], and Wakai *et al.* [53] did not recover the pan angles. We note that our method can even calibrate images in which trees line a street.

Off-the-shelf cameras. Following [53], we also evaluated calibration methods using off-the-shelf cameras to validate the effectiveness of our method. Figure 7 shows the qualitative results using off-the-shelf fisheye cameras using SL-MH for training. Our method meaningfully outperformed Lochman *et al.*'s method [32] in terms of recovered images. These results indicate robustness in our method for various types of camera projection.