Large-Scale Person Detection and Localization using Overhead Fisheye Cameras
Supplemental Material

Lu Yang[1*], Liulei Li[2*], Xueshi Xin[1] , Yifan Sun[3] , Qing Song[1] , Wenguan Wang[4†]
[1] Beijing University of Posts and Telecommunications [2] ReLER, AAII, University of Technology Sydney
[3] Baidu [4] ReLER, CCAI, Zhejiang University
https://LOAFisheye.github.io/

		# Video	# Image	# People			Scene			Season			Time		
				Total	Avg.	Max	Total	Indoor	Outdoor	Spring	Summer	Autumn	Morning	Noon	Afternoon
`train`		51	29,569	315,262	10.6	65	35	7	28	13	30	8	10	15	26
`val`	seen	3	1,700	18,460	10.8	29	3	1	2	1	2	0	1	1	1
	unseen	5	2,900	29,381	10.1	44	5	2	3	1	3	1	1	2	2
	total	8	4,600	47,841	10.4	44	8	3	5	2	5	1	2	3	3
`test`	seen	5	2,774	28,666	10.3	41	5	2	3	1	3	1	1	2	2
	unseen	10	5,999	65,993	10.0	44	10	2	8	3	5	2	3	3	4
	total	15	8,773	94,659	10.1	44	15	4	11	4	8	3	4	5	6
Total		74	42,942	457,762	10.5	65	50	11	39	17	38	11	14	20	32

Table S1: Detailed statistics of LOAF. # indicates the number of elements.

This document provides additional materials to supplement our main manuscript. We first present more statistics about LOAF in §A, and then give extra implementation details of our method in §B. More qualitative results on the `test` set of LOAF are summarized in §C. Next, we state the ethical conducts in §D. Finally, we provide the pseudo of our proposed rotation equivariant training strategy in §E.

A. Additional Dataset Analysis

More Statistics. LOAF is captured from multiple indoor/ outdoor scenes (*e.g.*, library, classroom, street, parking lot) across three seasons, we summarize the detailed statistics in Table S1, including the number of boxes, video sequences, *etc*. As seen, the majority of videos are collected from outdoor environments characterized by increased complexity, larger fields of view, and a higher number of human targets when compared to the indoor ones. These videos are divided into `train`, `val`, and `test` sets in the ratio of 7:1:2 respectively, while ensuring an roughly even distribution of attributes (*e.g.*, season, time) across these sets.

B. More Implementation Details

Training Objective. We extend the Generalized IoU (GIoU) loss [4] utilized in vanilla DETR [1] for bounding box regression to the rotated setup. Concretely, Brute-force search is leveraged to compute the minimum enclosing box between two rotated bounding boxes. It is implemented in a fully differentiable manner and adapted for parallel processing on GPU, which merely defers the training speed by around 5% when compared to the axis-aligned setup.

C. Qualitative Evaluation

Visual Comparison. Fig. S1-S5 compare our method with existing work qualitatively. It is obvious that our proposed method consistently presents more accurate detection and localization results, regardless of the category of scenes. Notably, it is much more effective than existing work for targets that are relatively small or densely arranged.

Diversity. To render a more intuitive understanding of the diversity of LOAF, a collage constituted from various scenes characterized by distinct attributes is given in Fig. S6.

D. Ethical Conducts

To protect the privacy of individuals and groups, we utilize Gaussian filters to blur all visible facial regions in LOAF. The proprietary data can only be accessed for non-commercial purposes to prevent inappropriate usage.

E. Pseudo Code

We offer the pseudo code for our proposed query-based rotation equivariant training strategy in Algorithm S1.

Algorithm S1 Pseudo-code for our proposed rotation equivariant training strategy.

```
"""
I: input image
gt: ground truth
angle: degree of clockwise rotation
$\lambda$: the balance factor
"""
def rotat_equi_training(I, gt):
    # $\mathcal{F}(I)$
    m1 = Encoder(I)
    angle = randint(0, 360)
    # $\mathcal{F}(g^r(I))$
    m2 = Encoder(rotate(I, angle))

    # $\{\boldsymbol{q}_n\}_{n=1}^N=\kappa(\boldsymbol{I})$
    query1 = gen_proposal(m1)
    # $\{\boldsymbol{q}_n^{g^r}\}_{n=1}^N=\kappa(\boldsymbol{I}^{g^r})$
    query2 = gen_proposal(m2)

    # $\mathcal{D}(\boldsymbol{I}, \{\boldsymbol{q}_n\}_{n=1}^N)$
    det1 = Decoder(m1, query1)
    # $\mathcal{D}(g^r(\boldsymbol{I}), \{\boldsymbol{q}_n^{g^r}\}_{n=1}^N)$
    det2 = Decoder(rotate(m1, angle), query2)

    # $\mathcal{L}_{\text{det}}\left(\{\boldsymbol{b}_{\ell(n)}\}_{n=1}^N, \mathcal{D}(\boldsymbol{I}, \{\boldsymbol{q}_n\}_{n=1}^N)\right)$
    loss1 = det_loss(det1, gt)
    # $\mathcal{L}_{\text{det}}\left(\{\boldsymbol{b}_{\ell(n)}^{g^r}\}_{n=1}^N, \mathcal{D}(g^r(\boldsymbol{I}), \{\boldsymbol{q}_n^{g^r}\}_{n=1}^N)\right)$
    loss2 = det_loss(det2, label_rotate(gt, angle)

    return loss1 + $\lambda$*loss2
```

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. S1

[2] Zhihao Duan, Ozan Tezcan, Hayato Nakamura, Prakash Ishwar, and Janusz Konrad. Rapid: rotation-aware people detection in overhead fisheye images. In *CVPR Workshop*, 2020. S3, S4, S5, S6, S7

[3] Shengye Li, M Ozan Tezcan, Prakash Ishwar, and Janusz Konrad. Supervised people counting using an overhead fisheye camera. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2019. S3, S4, S5, S6, S7

[4] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union. 2019. S1

[5] Roman Seidel, André Apitzsch, and Gangolf Hirtz. Improved person detection on omnidirectional images with non-maxima suppression. In *International Conference on Computer Vision Theory and Applications*, 2019. S3, S4, S5, S6, S7

[6] Masato Tamura, Shota Horiguchi, and Tomokazu Murakami. Omnidirectional pedestrian detection by rotation invariant training. In *WACV*, 2019. S3, S4, S5, S6, S7

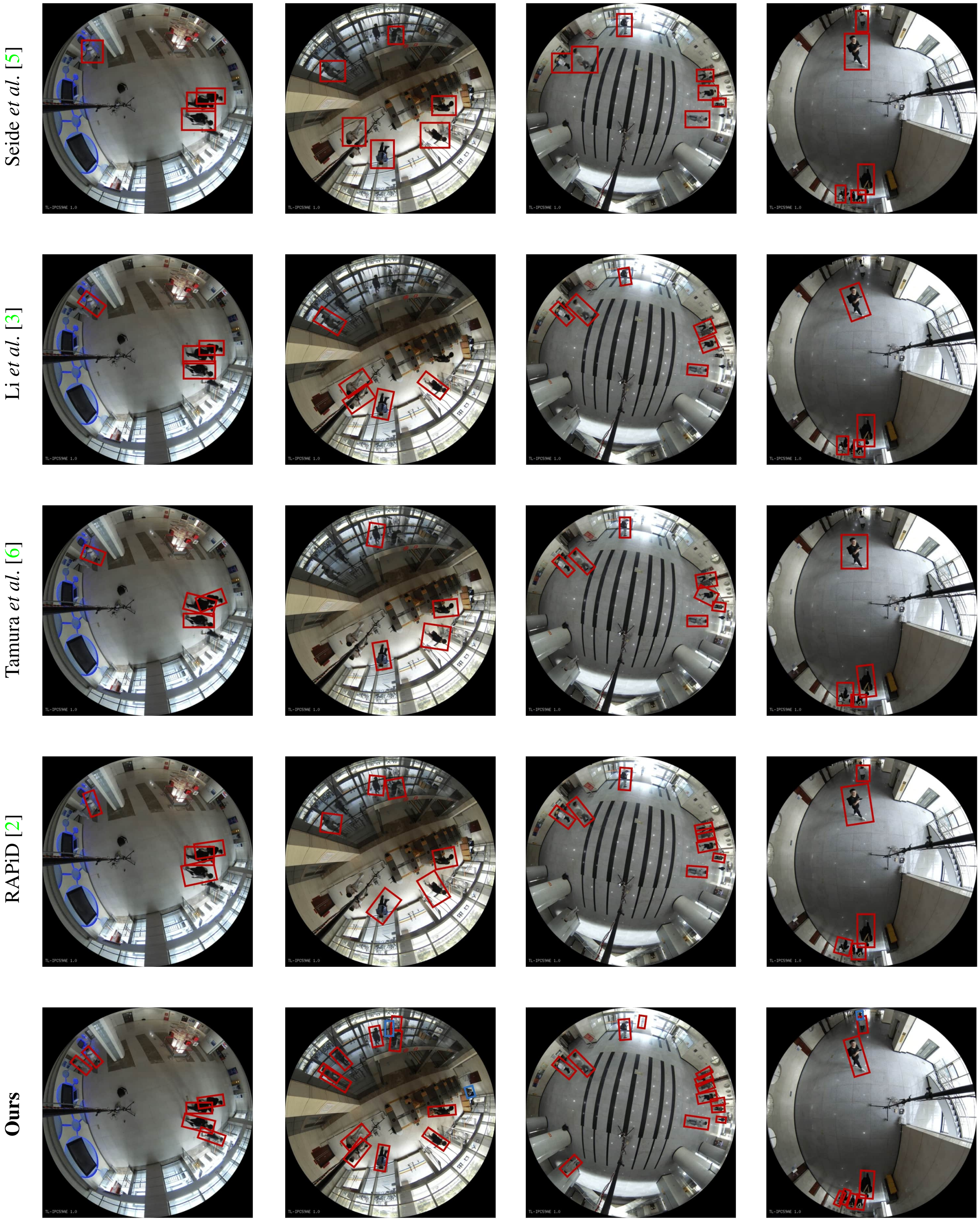

Figure S1: **Visual comparison of detection results** on the `test` set of LOAF. ◇ indicates targets missed by our method.

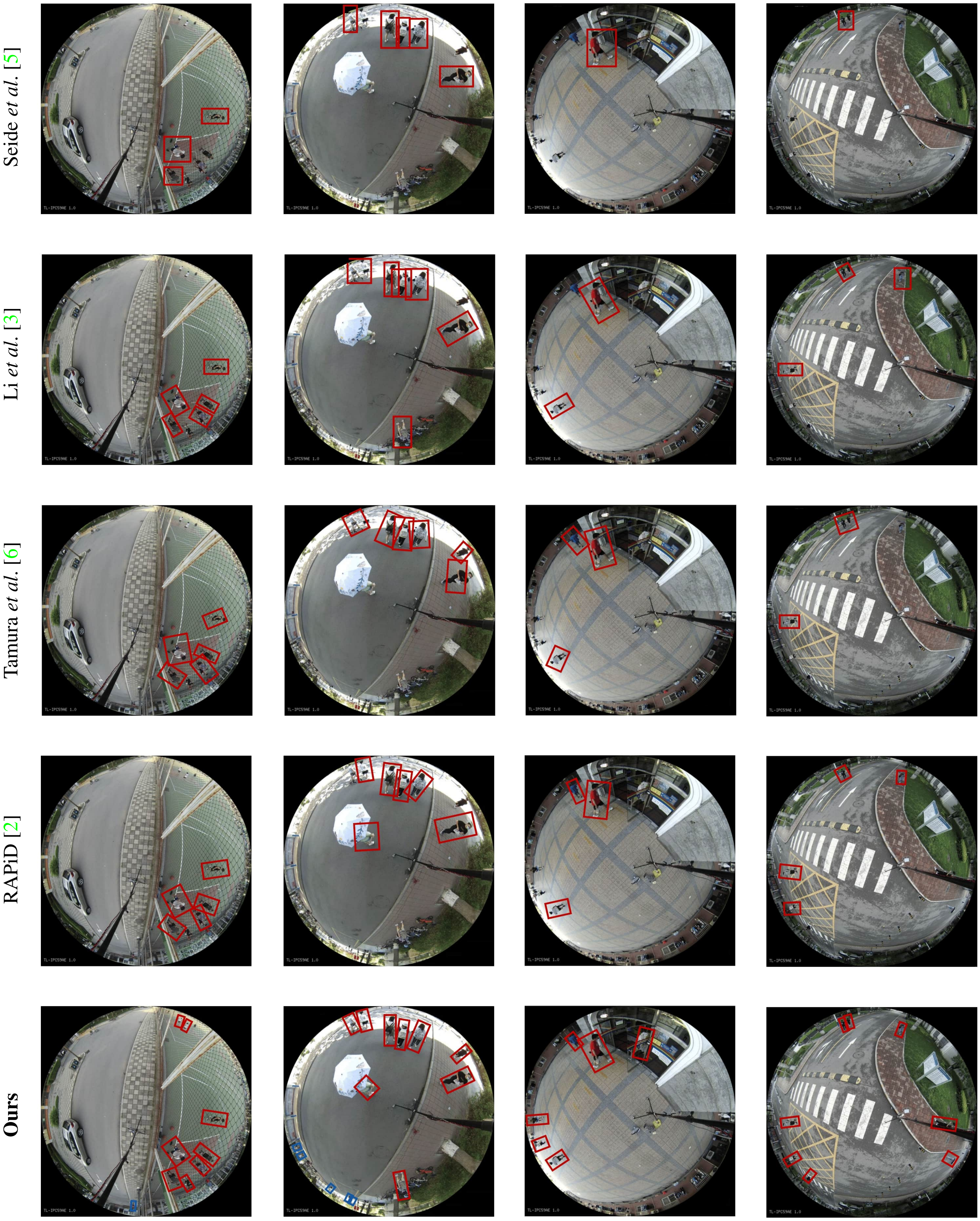

Figure S2: **Visual comparison of detection results** on the `test` set of LOAF. ◇ indicates targets missed by our method.

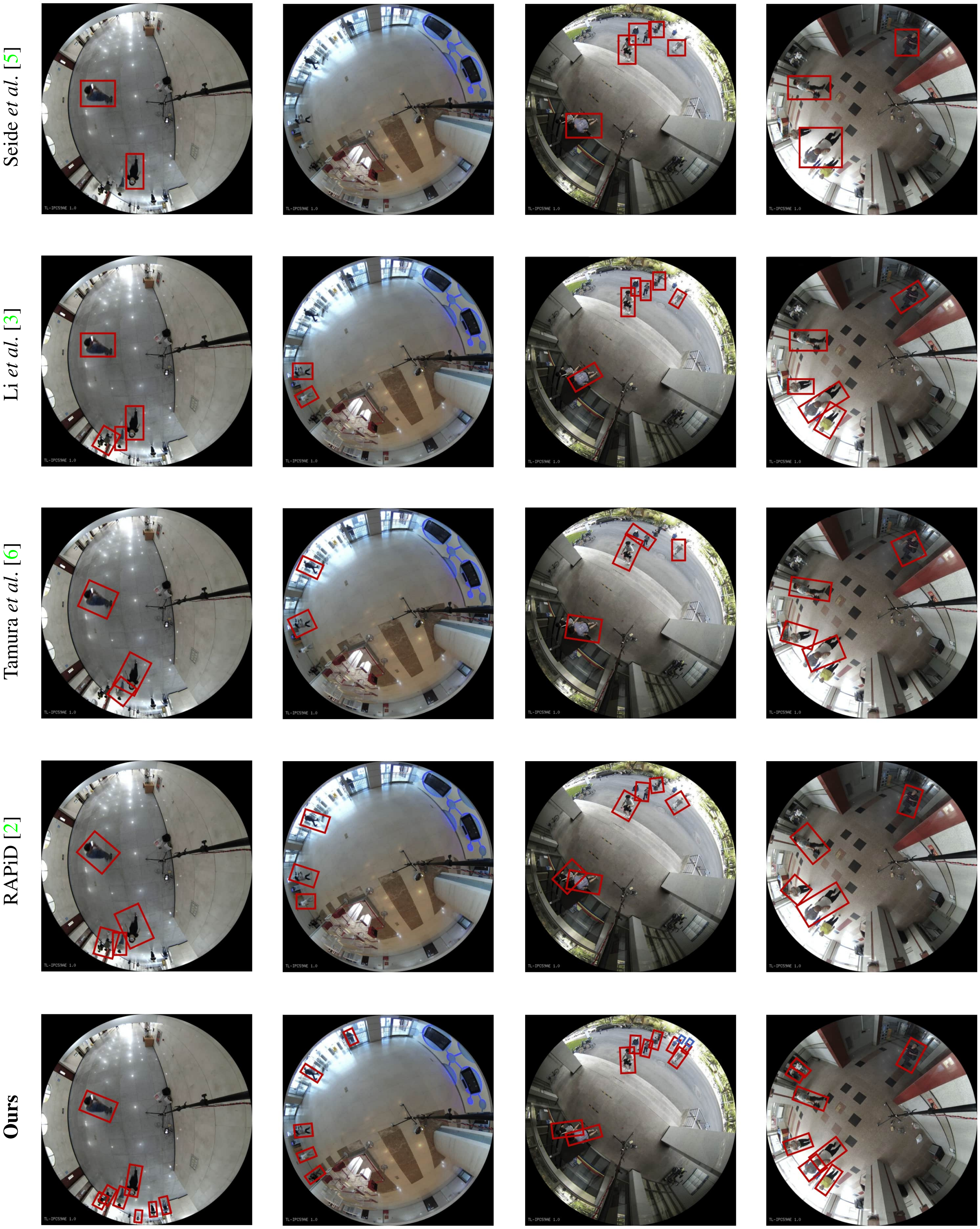

Figure S3: **Visual comparison of detection results** on the `test` set of LOAF. ◇ indicates targets missed by our method.

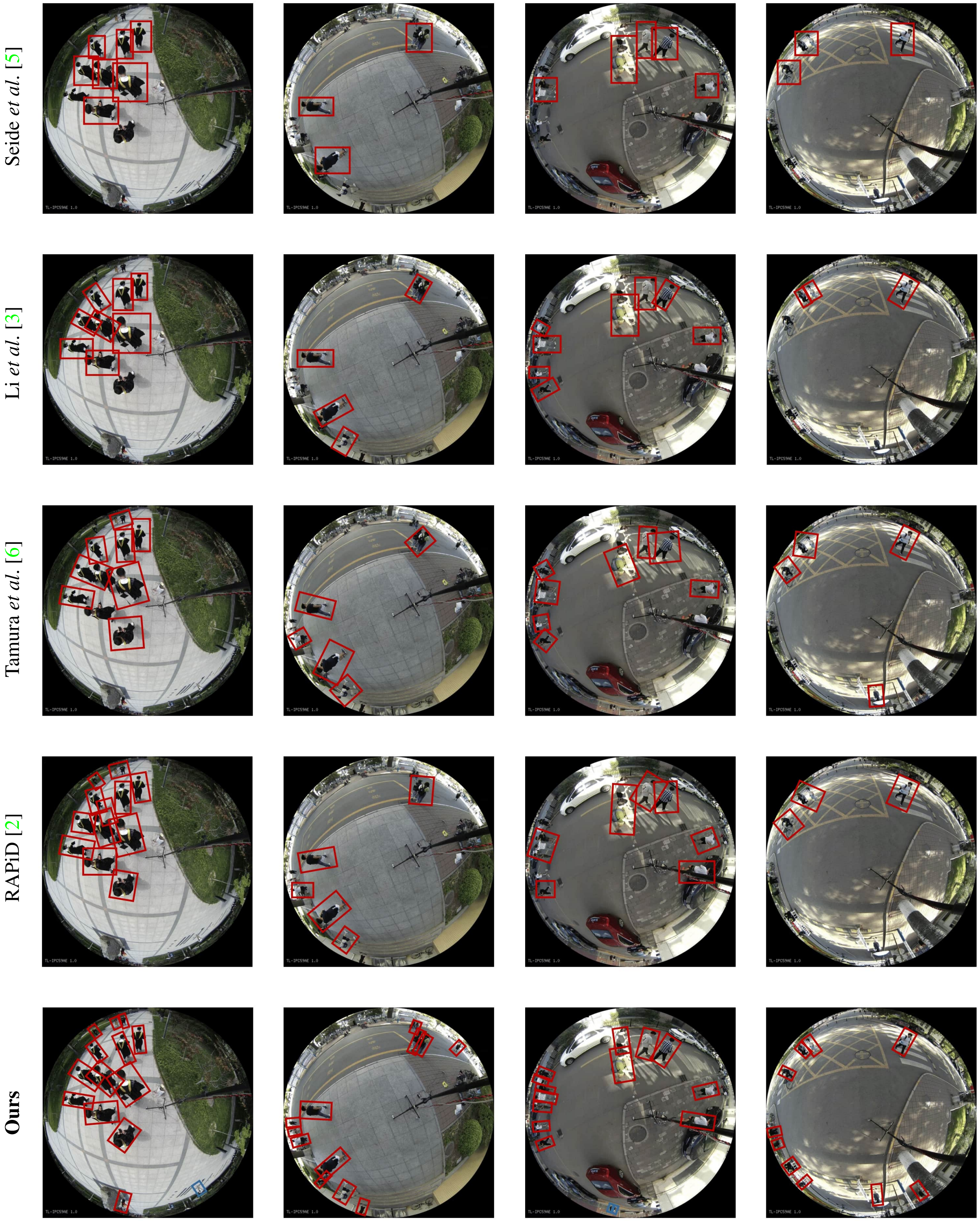

Figure S4: **Visual comparison of detection results** on the `test` set of LOAF. ◇ indicates targets missed by our method.

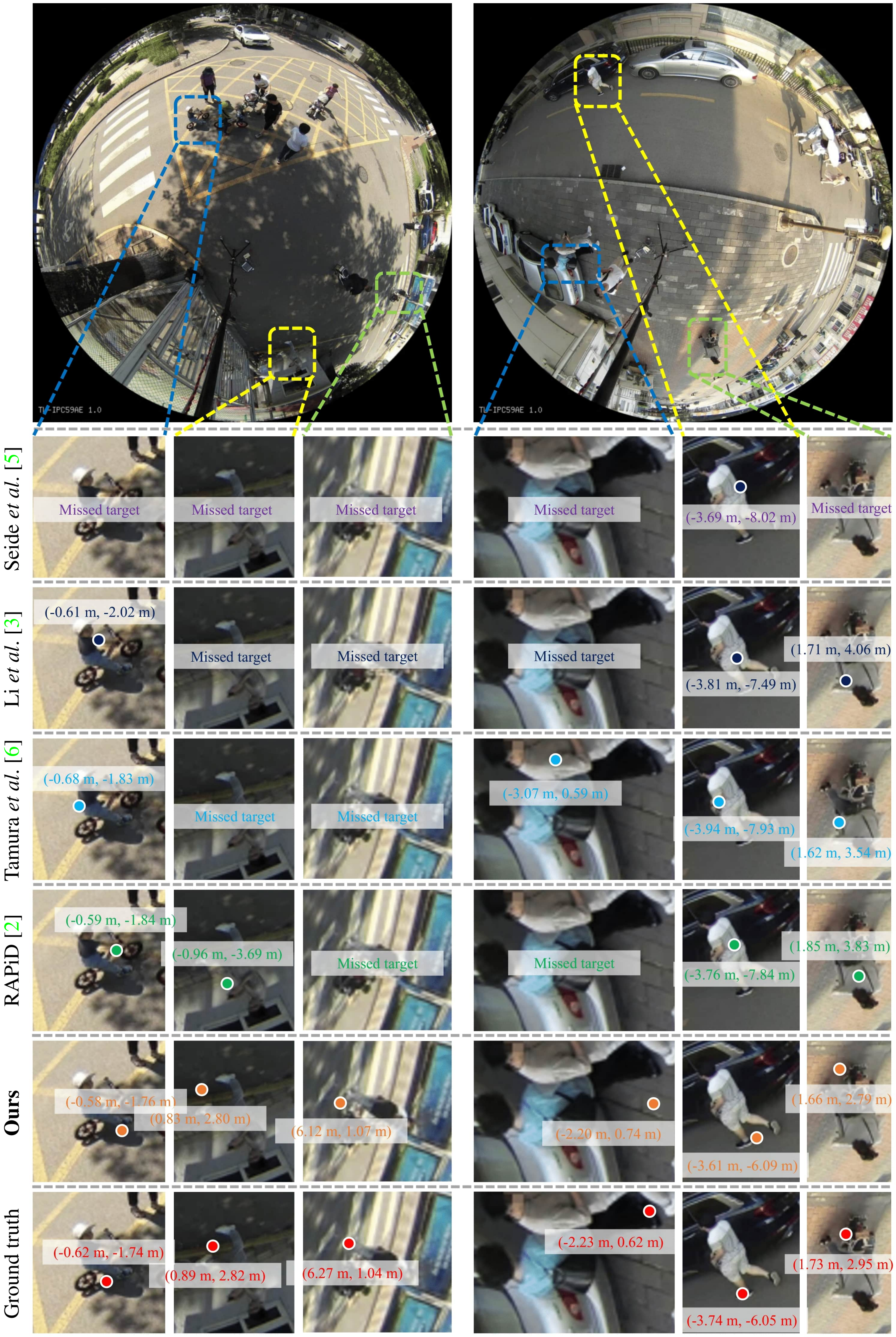

Figure S5: **Visual comparison of localization results** on the `test` set of LOAF. We selected three targets per frame for clear visualization.

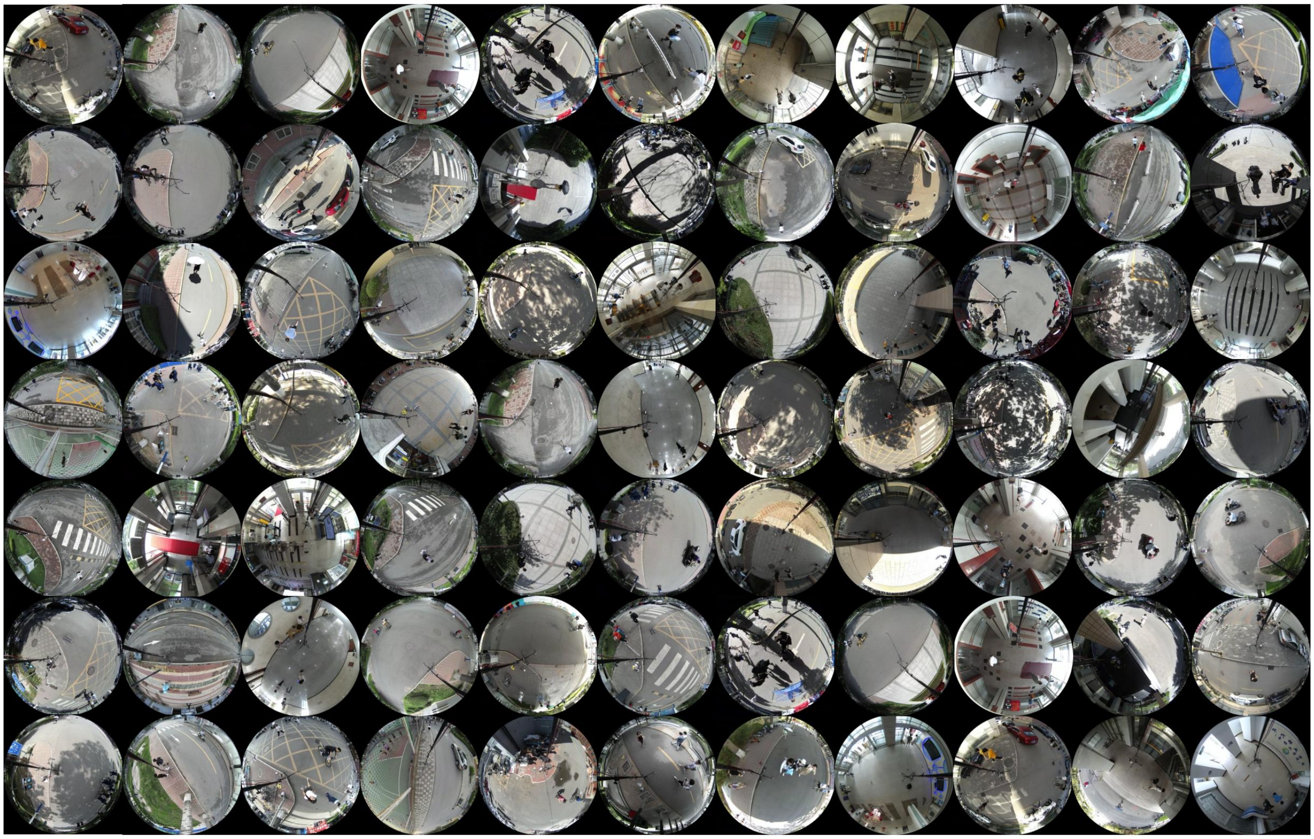

Figure S6: A collage constituted from various scenes characterized by distinct attributes.