

Large-Scale Person Detection and Localization using Overhead Fisheye Cameras

Lu Yang^{1*}, Liulei Li^{2*}, Xueshi Xin¹, Yifan Sun³, Qing Song¹, Wenguan Wang^{4†}

¹ Beijing University of Posts and Telecommunications ² ReLER, AAIL, University of Technology Sydney

³ Baidu ⁴ ReLER, CCAI, Zhejiang University

<https://LOAFisheye.github.io/>

Abstract

Location determination finds wide applications in daily life. Instead of existing efforts devoted to localizing tourist photos captured by perspective cameras, in this article, we focus on devising person positioning solutions using overhead fisheye cameras. Such solutions are advantageous in large field of view (FOV), low cost, anti-occlusion, and non-aggressive work mode (without the necessity of cameras carried by persons). However, related studies are quite scarce, due to the paucity of data. To stimulate research in this exciting area, we present LOAF, the first large-scale overhead fisheye dataset for person detection and localization. LOAF is built with many essential features, e.g., i) the data cover abundant diversities in scenes, human pose, density, and location; ii) it contains currently the largest number of annotated pedestrian, i.e., 457K bounding boxes with ground-truth location information; iii) the body-boxes are labeled as radius-aligned so as to fully address the positioning challenge. To approach localization, we build a fisheye person detection network, which exploits the fisheye distortions by a rotation-equivariant training strategy and predict radius-aligned human boxes end-to-end. Then, the actual locations of the detected persons are calculated by a numerical solution on the fisheye model and camera altitude data. Extensive experiments on LOAF validate the superiority of our fisheye detector w.r.t. previous methods, and show that our whole fisheye positioning solution is able to locate all persons in FOV with an accuracy of 0.5 m, within 0.1 s.

1. Introduction

Accurate position finding of persons attracts growing interest from both research and industrial communities, since it plays a crucial role in numerous location-sensitive application scenarios (e.g., surveillance, smart home, public health). Nevertheless, due to the line-of-sight (LOS) issue, GPS is unreliable in interior spaces and urban canyon. To overcome

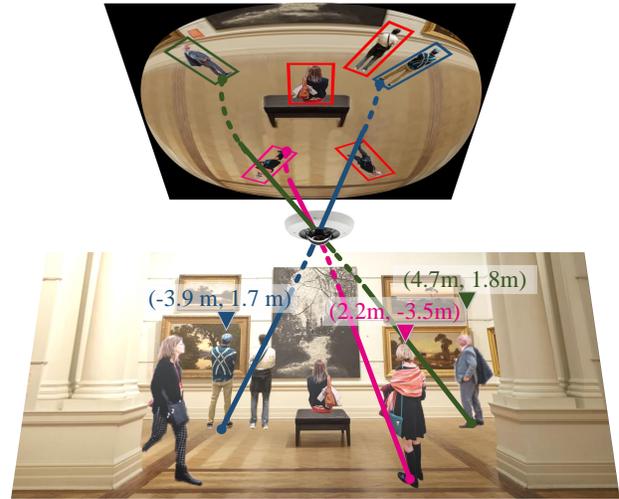


Figure 1: **Person positioning using overhead fisheye camera:** We detect humans on omnidirectional images and then project detects onto the real world coordinates to obtain physical locations. Compared with using perspective cameras, our fisheye camera based solution is favored in low cost, high accuracy, and fast speed.

such limitation, various alternative solutions are investigated. *Signal based* solutions, including Bluetooth [13] and Wi-Fi [74], are popular, but they are easily interfered by changing environments and nearby human bodies [73]. A complementary stream of work is *vision based*; they typically make use of traditional cameras, RGBD cameras, or in-built smartphone cameras, and enjoy the advantage of reliable services. To get location information, visual positioning solutions usually refer to a pre-acquired 3D map or a geo-tagged database as the scene representation [63], or directly utilize the captured image to estimate the camera pose [38].

Although visual localization has been a hotspot issue for many years, existing efforts are mainly dedicated to urban place recognition or indoor camera localization, based on *perspective cameras* [63, 38, 62, 42]. None of them addresses person positioning by using *overhead fisheye cameras*, even though fisheye cameras are widely used in visual surveillance applications. One possible reason is the lack of accessible datasets, compounded by considerable costs in-

¹The first two authors contribute equally to this work.

²Corresponding author: Wenguan Wang.

volved in data collection. In this article, we provide a large-scale overhead fisheye dataset, LOAF, for person detection and localization in both indoor and outdoor scenes.

Compared with perspective cameras, overhead (top-view) fisheye cameras are promoted due to less occlusion among people and larger field of view (FOV) – allowing the coverage of a large space using a single, low-cost camera. Only few public datasets [21, 41, 24] provide top-view fisheye data. Unfortunately, they only cover quite few scenes and their people-box annotations do not adequately address the positioning challenge, negatively affecting location approximation (see §3.2 for detailed analysis). Differently, LOAF specifically targets at person localization in surveillance applications, and has the following appealing characteristics:

- *Large-scale*: LOAF is the largest in the field, to our best knowledge. It consists of over 70 videos, with more than 43K frame images, 457K person-detection annotations as well as corresponding location information.
- *High diversity*: LOAF contains a wide variety of surveillance scenarios; it includes a total of 11 indoor and 40 outdoor scenes, and the data are captured at different times of day and cover different illumination conditions.
- *Positioning-aware person-box annotation*: LOAF offers radius-aligned human-box annotations. Compared with other person representations, *i.e.*, head center [21], axis-aligned [60] or body-aligned box [41, 24], used in previous fisheye datasets, the radius-aligned box is promoted, as it fits well radially-oriented bodies [41], and, more importantly, it is better aware of the positioning problem.

Moreover, we devise a person positioning system that first detects persons from raw fisheye images, and then calculates physical locations based on fisheye visual model and altitude information (see Fig. 1). Clearly, high-quality person detection is the crucial premise for precise localization using overhead fisheye cameras. Fisheye lenses provide large FOV, at the cost of strong radial distortion. This makes pedestrian detection in top-view fisheye images a much harder task, compared with using perspective cameras. Studies on overhead fisheye pedestrian detection are very scarce [60, 41, 50, 24, 64]; they rarely concern the link with person positioning, and many of them [60, 41] are even not trainable due to the lack of fisheye data. With our LOAF dataset, we develop a novel query based fisheye human detector. It explicitly exploits fisheye geometry by accommodating *rotation equivariance* into the matching between queries and human instances during network training. The insight here is intuitive: for a robust fisheye detector, rotation of a fisheye image should result in correspondingly rotated detections. In addition, our detection algorithm learns to predict radius-aligned person boxes, facilitating localization estimation.

We test our fisheye person detection algorithm as well as our whole positioning system over LOAF. We find that our detector significantly outperforms previous methods, and

our full system delivers precise localization results. We also empirically show that our algorithm generalizes well on previous top-view fisheye person detection datasets [24, 65].

2. Related Work

Accurate Positioning. GPS is the most popular system for outdoor localization. As it requires LOS between the satellites and the handset, GPS does not function well in urban canyons, indoors and basements [30, 26]. This triggered the development of alternative positioning solutions, following a multi-disciplinary approach. Concretely, there are two main schemes of the alternatives: *signal-based* and *vision-based*.

Signal-based positioning systems typically lean on sound wave [49], geomagnetism [31], radio frequency (RF) [26, 74, 54, 13, 1], and infrared radiation (IR) [71], as well as different location determination techniques, such as TOA (time of arrival) [30] and RSS (received signal strength) [26]. The main challenge to signal-based systems is the sensitivity to environment changes, such as object moving, diffraction and reflection, which affect signal propagation [26].

Visual data is another potential information source for precise localization. Since put forward [61], visual positioning has become a hot topic in robotics and computer vision. Some methods adopt a pre-built geo-tagged database or a 3D scene model, as the reference for camera location estimation [32, 77, 66, 78, 76]. Some others rely on recognizing some deployed coded targets [27, 59, 20, 35], *e.g.*, concentric rings, barcodes, colored dots, *etc.* Some recent ones utilize deep learning techniques to replace some components (*e.g.*, image retrieval, descriptor matching) in traditional systems [4, 16], or regress the camera pose directly [38, 37, 7]. Though promising, existing visual positioning systems are mostly founded on perspective cameras.

Due to the problem complexity, there is no persuasive solution for precise positioning yet, and hybrid schemes are often applied in practice. Fisheye cameras have advantage of providing wide FOV with low cost and reduced occlusion, while the research landscape is sparse for fisheye camera based localization [79, 21]. To foster research in this direction and facilitate practical deployment, we contribute the first overhead fisheye dataset, to our knowledge, that allows to conduct the task of person localization at large scale.

Deep Learning based Visual (Camera) Localization. Regarding application scenarios, scholars in computer vision community are mainly aware of *city-scale location recognition* [2, 72, 58, 57] and *indoor camera localization* [5, 67]. Popular visual localization approaches can be divided as *retrieval-based*, *regression-based*, and *structure-based*, according to the camera pose estimation strategy. Retrieval-based methods [58, 3, 23, 75, 56] represent a scene as a database of geo-tagged photos and use geo-tag of the most relevant database photo as an approximation to the camera position [58]. Regression-based methods learn to encode

the scene into a deep network and directly regress a 6DOF camera pose [38, 67, 37] from a captured image. Structure-based methods [6, 55, 29] pose the localization problem as a camera resectioning task [58]. They first represent scenes via 3D models and establish a set of 2D-3D matches, then recover the full camera pose by employing a PnP solver [33] inside a RANSAC [28] loop.

These visual positioning techniques seek to localize images captured by handheld devices [42, 11] or vehicle cameras [47] and demand pre-created geo-tagged databases or 3D maps. And they work on an *active* mode – users need to carry the cameras. Differently, we address person localization in surveillance scenarios, by using stationary, overhead fisheye cameras. This yields a *passive*, low-cost scheme, where the localization is completed by a numerical solution on the fisheye model fused with altitude information. Thus our scheme faces a different challenge, *i.e.*, conducting robust person detection from non-rectilinear fisheye images. Overall, our scheme fills the gap left by conventional studies and is complementary to existing positioning systems.

Person Detection in Overhead Fisheye Images. As a canonical subproblem of object detection, pedestrian detection has long received great interest owing to its broad applications such as intelligent surveillance and autonomous vehicles [12, 8]. By contrast, people detection in overhead fish-eye images has been studied much less, due to the absence of such datasets. Moreover, adapting standard pedestrian detectors to top-view fisheye cameras is difficult [15]: First, the appearance of people in fisheye images is arbitrary-oriented, while in perspective images, people typically appear upright. Second, people suffer from severe geometric distortions, particularly in the fisheye image’s periphery.

Faced with the above challenges, early attempts make use of handcrafted features (*e.g.*, HOG, LBP) as well as standard pedestrian detectors with slight modifications to account for fisheye geometry [39, 14, 17, 69]. The common approach is dewarping fisheye images [14] or features [39] so as to approximate normal people’s appearances from the deformed ones. Though this simplifies the classification of features, approximation errors are inevitable, causing performance degradation [64]. Later, a few CNN-based fisheye detectors were developed. Some of them are *training-free*. For example, [60] runs standard YOLOv2 [51] on dewarped versions of overlapped windows extracted from a fisheye image. Li *et al.* [41] rotate a fisheye image in 15° increments, and apply off-the-shelf YOLOv3 [52] only to the top-center region of each rotated image, where people usually appear upright. Some more recent methods are *trainable*: [64] trains YOLOv2 with rotated perspective images so as to handle omnidirectional images without test-time transformation; [24] trains YOLOv3 with human-aligned boxes.

Although many prior arts [60, 41, 64] allowing to detect persons without any fisheye training data, they require a cer-

tain amount of computation time for post-processing: [60] carries out detection in multiple perspective images dewarped from one omnidirectional image; [41] applies YOLOv3 24 times to each fisheye image; [64] needs a grouping process to eliminate numerous redundant results, caused by a rotation-invariant training strategy. Hence, the utility of previous methods is limited in the context of visual positioning. Our analysis in §3.2 sheds light on the weaknesses of human representations adopted by existing fisheye detectors (*e.g.*, head center [21], horizontal or body-aligned box [60, 41, 24]) in regard to localization. Thus we supply our large-scale dataset with positioning-aware person-detection annotation. As a result, our human detector can benefit from end-to-end, fisheye visual pattern learning and output radius-aligned body-boxes for precise positioning. Further, by regularizing the learning of image representations and instance queries with rotation equivariance, our algorithm naturally addresses the omnidirectional nature of fisheye images, yet using standard detection network architecture designed for perspective images.

3. LOAF Dataset

3.1. Dataset Acquisition

We first describe how our fisheye images are collected. **Apparatus and Technical Specifications.** A TL-IPC59AE fisheye camera with 1.1mm focal length is adopted for data recording. It has a wide FOV, reaching the full circle in the horizontal plane and 180° in the vertical plane. This offers a clear advantage in reducing deployment cost – installing just one fisheye camera instead of multiple conventional cameras to monitor the same region. However, the severe geometric distortions introduced prevent the use of standard detectors, which are designed for conventional cameras [8]. **Data Capturing.** The fisheye camera is mounted on the ceiling (indoors) or poles (outdoors), 2.5~4.0 m from the ground with 200~300 m² FOV. Considering the factors such as scenario diversity, pedestrians density and weather condition, we capture 110 fisheye image sequences in 80 realistic scenarios as the raw data pool. The recorded sequences span 14 hrs; the image resolution is 2952×2952 pixels, and the frame rate is 10~20 fps. Eventually, 42,942 images, sampled at 1 fps, are collected to construct our LOAF dataset.

3.2. Dataset Annotation

We next describe how our fisheye images are annotated. **Person Detection.** As the performance of localization relies critically on the quality of person detections, significant effort should be spent on the fisheye person-detection annotation. Some human-detection representations were explored in previous datasets: [21] labels the *center point* of each human head; [41, 24] opt for *human-aligned* person-boxes. However, these representations have several shortcomings, especially with regards to the localization task. First, they suffer from some inherent limitations. The point-based rep-

Dataset	#Scene		#Video	#Image	#People			Max Resolution	FOV (m ²)	Annotation			FPS
	Indoor	Outdoor			Total	Avg.	Max			People Detection	Location	Attribute	
PIROPO [21]	2	0	27	-	-	-	-	800×600	<50	Head Center			10
HABBOF [41]	2	0	4	5,837	20,466	3.5	5	2,048×2,048	<36	Human-aligned box			12~30
MW-R [24]	6	0	19	8,752	22,825	2.6	6	1,488×1,488	<36	Human-aligned box			15
CEPDOF [24]	5	0	8	25,504	173,073	6.8	13	2,048×2,048	<36	Human-aligned box			1~10
WEPDOF [65]	14	0	16	10,544	93,363	8.9	35	2,592×1,944	<36	Human-aligned box			1~10
LOAF	11	39	74	42,942	457,762	10.5	65	2,952×2,952	200~300	Radius-aligned box	✓	✓	10~20

Table 1: **Overhead fisheye datasets comparison** (§3.4). LOAF is the largest in terms of the total number of pedestrian and scene categories.

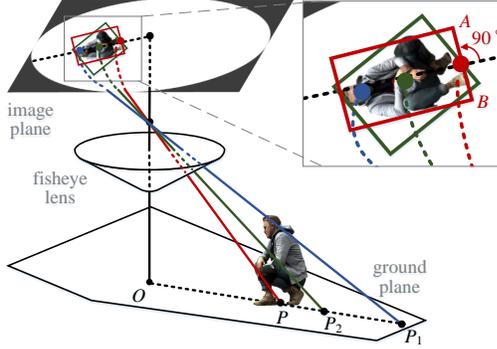


Figure 2: **Human annotation comparison for visual positioning** (§3.2). Previous methods typically use the detected human head [21] (●) or the center (●) of human-aligned body-box (◇) [41, 24] to determine the physical location P , bringing positioning errors (i.e., PP_1 , PP_2). We instead leverage the radius-aligned box (□) as human representation. The midpoint (●) of side AB , i.e., the closest point to the image center on the radius-aligned box, better corresponds to the human actual position.

representation is not applicable in other analysis tasks (e.g., re-identification, attribute recognition); the human-aligned box has ambiguity [10, 68] – the ground-truths of human-aligned boxes are not uniquely determined [24]. Second, and most important, these human-detection representations cannot meet the need of precise localization. Clearly, based on the fisheye camera model, one can project 2D detections on the 3D world for localization. However, the center of human head or of human-aligned box is not the exact placement occupied by human on the image plane, causing errors to estimations of human physical position (see Fig. 2).

We instead label each person through a *radius-aligned* rectangular box. Such representation is favored as it: i) allows unique groundtruth box assignment; ii) fits well radially-oriented human bodies presented in fisheye images; and iii) better corresponds to the actual position of human on the image plane, facilitating physical localization. Notably, although the radially-oriented box constraint was explored in a prior fisheye detector [41], there is no previous datasets provide such kind of annotation, neither is there any literature points out the advantage of such representation in localization. Finally, around 457K human box annotations are obtained. High-quality annotation is ensured via a rigorous quality check, conducted by highly skilled reviewers.

Person Localization. For each scene, a 10 m ruler (with 0.05 m accuracy) is placed on the ground, and one end of the

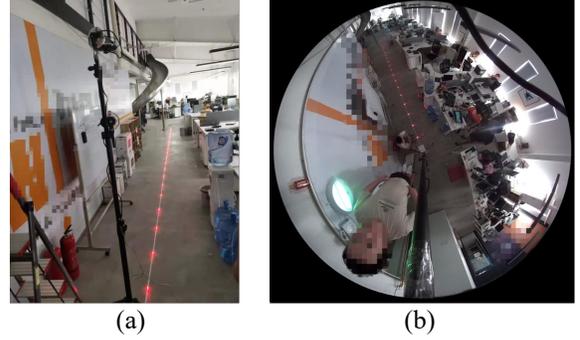


Figure 3: **Calibration for groundtruth person location annotation.** (a) Ground makers (with 0.05m precision). (b) Top-view fisheye image capture for calibration. See §3.2 for details.

ruler is directly under the fisheye lens (cf. Fig. 3 (a)). We take a fisheye picture and use this marked picture as the calibration for all the dataset images recorded in this scene (cf. Fig. 3 (b)). Finally, for each annotated human, the physical location, at sub-decimeter precision, is provided (cf. Fig. 4). **Scene Attribute.** To enable in-depth analysis, each image is annotated with multiple attributes, including day/night, outdoor/indoor, and sunny/rain/foggy/snow.

3.3. Dataset Design

We then list several key aspects of our dataset design.

Privacy Protection. To protect personal information, we apply the gaussian filter to blur all the visible facial regions in our dataset, and conduct experiments on the blurred data.

Dataset Splits. LOAF contains 29,569 training, 4,600 validation, and 8,773 testing images (approximately 7 train, 1 val, and 2 test). Moreover, to better evaluate models' generalization ability, LOAF is split into five sets: train (7/28 indoor/outdoor scenes, 29,569 images), val seen (1/2 indoor/outdoor scenes, 1,700 images), val unseen (2/3 indoor/outdoor scenes, 2,900 images), test seen (2/3 indoor/outdoor scenes, 2,774 images), test unseen (2/8 indoor/outdoor scenes, 5,999 images). There are no overlapping scenes between unseen and train sets.

Dataset Accessibility. Only the desensitized version of our dataset will be released online, under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License [19].

3.4. Dataset Features and Analysis

Finally, we present statistic analysis of LOAF in comparison with existing overhead fisheye datasets [21, 41, 24].

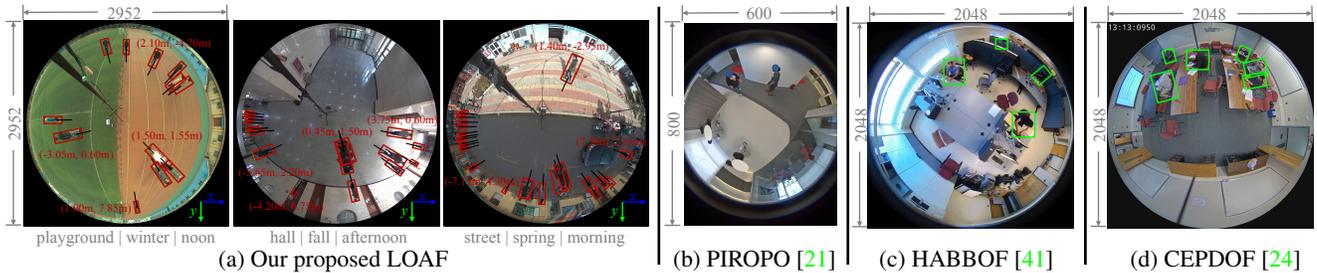


Figure 4: **Example images** from different datasets (§3.2-§3.4). Prior datasets [21, 41, 24] are restricted to few indoor scenes with person detection annotation only, *i.e.*, head center (●) or human-aligned box (□). In contrast, LOAF covers challenging indoor and outdoor scenes with human detection, localization and scene attribute annotations. The person-detection annotation is given as the radius-aligned box (▭), which is more suitable for localization. For better visualization, we only present location ground-truths for some of persons.

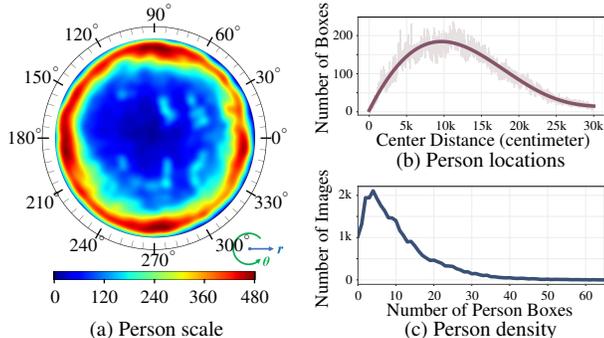


Figure 5: **Dataset Statistics** (§3.4): We summarize LOAF with the distribution of (a) person scale (area in pixel), (b) person locations (horizontal distance between human and the fisheye lens in the real world), and (c) person density (number of person per image).

LOAF distinguishes itself from three aspects (*cf.* Table 1):

Large-scale: LOAF has 42,942 fisheye images with more than 457K person boxes. Moreover, LOAF data are captured by an advanced fisheye camera, which is capable of covering a larger area (200~300 m²) with higher pedestrian density (2~65 persons per scene, 10.5 in average). This makes LOAF the largest overhead fisheye dataset in terms of the total number of pedestrian and scene categories.

High Diversity: Existing datasets limit in data diversities, *i.e.*, only containing very few indoor scenes (2~14) and completely missing outdoor scenarios (*cf.* Fig. 4). In contrast, LOAF involves 51 realistic scenes, including 11 indoor scenes (*e.g.*, lab, office, library, classroom) and 40 outdoor scenes (*e.g.*, street, playground, parking lot, square). The recorded data cover four seasons under different illumination (*e.g.*, morning, noon, afternoon) and weather (*e.g.*, sunny, rain, snow) conditions, and involve vast variance of human pose (*e.g.*, walking, standing, and sitting), scale, location, and density (*cf.* Fig. 5). Thus our dataset better reflects the distribution in real-world surveillance scenarios.

Rich and Positioning-aware Annotation: LOAF is provided with rich ground-truths for detection, localization, and scene attribute, which lays a solid foundation for fisheye camera based human-centric analysis. Hence, as demonstrated in §3.2, the radius-aligned human-box representation is adopted during our annotation. Compared with human-

head center based point annotation [21] and human-aligned person-boxes [41, 24] used in previous datasets, radius-aligned human-boxes are more suitable for the position task.

4. Our Approach

Our overhead fisheye camera based person localization solution comprises two parts. The former computes 2D detections that locate the people on the image plane (§4.1). The latter converts the 2D detections to 3D-world coordinates, obtaining the physical location of the people (§4.2).

4.1. Overhead Fisheye Person Detection

Core Idea: Rotation Equivariance. One of the reasons for the tremendous success of CNNs is their *equivariance* to horizontal and vertical shifts and the resulting invariance to local deformations [40, 25]. This stimulates a line of efforts to learn robust representations equivariant to generic types of transformations [34, 18, 70]. Formally, a representation f is said to be equivariant with a geometric transformation g (*e.g.*, cropping, flipping) for an input (image) I if:

$$f(g(I)) \approx g(f(I)). \quad (1)$$

That is to say, the output representation $f(I)$ is changed in the same way as the transformation g imposed to the input I .

In this work, we devise a *query*-based fisheye person detector that exhibits 360°-rotational equivariance through an elaborately designed training protocol. Our key insight is derived from the omnidirectional nature of this task: if one rotates the input fisheye image by an arbitrary angle, then the outputs of a robust fisheye detector should change accordingly. We thus design a *rotation equivariant* training strategy, which forces the matching between object queries and image representations to be equivariant against 360°-rotations. In this way, the rotational symmetry of omnidirectional images is explicitly addressed, without architectural modification of standard detectors developed for perspective images.

Rotation Equivariant Training for Query-based Fisheye Detection. As shown in Fig. 6, our fisheye detector is built upon DETR [9], promoted by a rotation equivariant training strategy. Basically, DETR conducts detection in a query-based fashion. Denote \mathcal{F} as a *feature encoder* (blue box) that extracts representation I of image I , and κ as a *query creator* (yellow box)

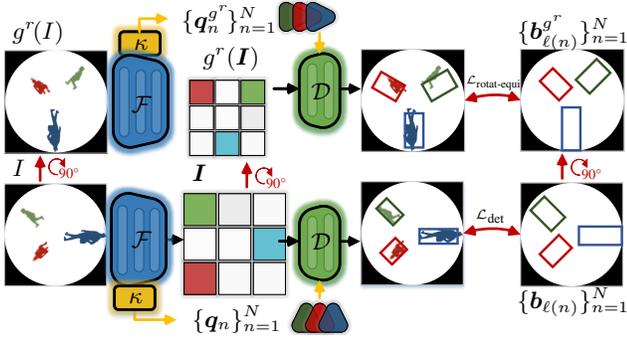


Figure 6: **Our rotation equivariant training strategy** for query-based fisheye person detection (§4.1). For equivariant training, object queries $\{\mathbf{q}_n^{g^r}\}_{n=1}^N$ created for rotated image $g^r(I)$ and the rotated representation $g^r(I)$ are fed into the decoder \mathcal{D} ; the rotated ground-truths $\{\mathbf{b}_{\ell(n)}^{g^r}\}_{n=1}^N$ are set as the training targets.

that outputs a set of N object-aware descriptors $\{\mathbf{q}_n\}_{n=1}^N$ from I . DETR employs $\{\mathbf{q}_n\}_{n=1}^N$ as queries to retrieve target objects from I :

$$\{\hat{\mathbf{b}}_n\}_{n=1}^N = \mathcal{D}(I, \{\mathbf{q}_n\}_{n=1}^N), \quad (2)$$

where $I = \mathcal{F}(I)$, $\{\mathbf{q}_n\}_{n=1}^N = \kappa(I)$, and \mathcal{D} refers to the Transformer *decoder* which formulates the object-query matching process via neural cross-attention computation. $\{\hat{\mathbf{b}}_n\}_{n=1}^N$ are the set of predicted parameterized object bounding boxes.

Our rotation equivariant training further encourages robust object-query matching which is equivariant against rotations. First, given an input rotation transformation g^r , a robust fisheye detector is desired to be able to extract rotation-equivariant representation. Analogous to Eq. 1, we have:

$$\mathcal{F}(g^r(I)) \approx g(\mathcal{F}(I)) = g^r(I). \quad (3)$$

Also, it is reasonable to assume that the final output of a robust fisheye detector should be changed in the same way to the rotation g^r applied to the input fisheye image I :

$$\{\mathbf{b}_{\ell(n)}^{g^r}\}_{n=1}^N \approx \mathcal{D}(I^{g^r}, \{\mathbf{q}_n^{g^r}\}_{n=1}^N), \quad (4)$$

where I^{g^r} indicates the feature of rotated image $g^r(I)$, i.e., $I^{g^r} = \mathcal{F}(g^r(I))$. Similarly, $\{\mathbf{q}_n^{g^r}\}_{n=1}^N$ denote the object queries derived from $g^r(I)$, i.e., $\{\mathbf{q}_n^{g^r}\}_{n=1}^N = \kappa(I^{g^r})$. $\{\mathbf{b}_{\ell(n)}^{g^r}\}_{n=1}^N$ are the rotated groundtruth bounding boxes, where $\ell(n)$ returns the groundtruth index for n -th query.

Considering Eq. 3 and Eq. 4, we can have: $\{\mathbf{b}_{\ell(n)}^{g^r}\}_{n=1}^N \approx \mathcal{D}(g^r(I), \{\mathbf{q}_n^{g^r}\}_{n=1}^N)$. Hence our rotation equivariant training objective is given as:

$$\mathcal{L}_{\text{rotat-equiv}} = \mathcal{L}_{\text{det}}(\{\mathbf{b}_{\ell(n)}^{g^r}\}_{n=1}^N, \mathcal{D}(g^r(I), \{\mathbf{q}_n^{g^r}\}_{n=1}^N)). \quad (5)$$

Here \mathcal{L}_{det} is the standard detection loss in DETR [9]. As such, the rotation equivariance properties for both fisheye representation \mathcal{F} (cf. Eq. 3) and the query-based detection prediction $\mathcal{D}(I, \{\mathbf{q}_n\}_{n=1}^N)$ (cf. Eq. 4) are sought in a single training target. This also allows us to effortlessly adapt standard

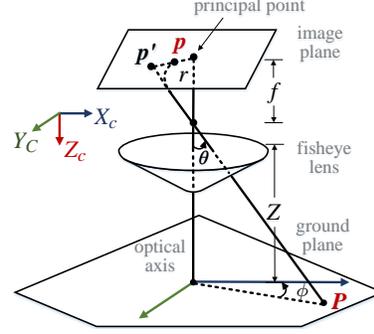


Figure 7: **Our fisheye camera model based person localization** (§4.2). Here p and p' denote the projection point of the distortion point P by fisheye lens and normal perspective lens, respectively.

DETR to our omnidirectional detection setting. Note that our rotation equivariant training differs from rotational data augmentation technique which only views rotated images as individual training samples (see §5.4 for detailed experiments).

4.2. 2D-3D Projection based Person Localization

General Fisheye Camera Model. The perspective projection of a normal pinhole camera can be written as $r = f \tan \theta$, where r indicates the projection distance between the principal point and the image point, f is the focal length, and θ is the angle between the incident ray and the camera’s optical axis. However, fisheye lens does not follow this perspective projection model [48], as the FOV equals to 180° (cf. Fig. 7). Fortunately, the image formation of different kinds of fisheye lenses can be approximated by a general polynomial projection model [36], i.e.,

$$r(\theta) = \sum_{i=1}^n k_i \theta^{2i-1}, \quad n = 1, 2, 3, 4, \dots \quad (6)$$

High distortions can be handled well when $n = 5$ [36]. The coefficients k s can be obtained from camera calibration.

2D-3D Projection. Given a detected human location point $p = (u, v)^\top$ in the fisheye image pixel coordinate system, our target is to calculate its 3D location $P = (X, Y, Z)^\top$ in the camera coordinate system, where Z is the altitude of the fisheye camera and priorly known. To do so, one can first calculate r as: $r = \sqrt{(x^2 + y^2)}$, where $x = (u - u_0)/f$ and $y = (v - v_0)/f$, and $(u_0, v_0)^\top$ are the coordinates of the principal point in the fisheye image. Then θ can be obtained by solving Eq. 6 using a numerical means [36]. With the altitude Z of the camera, the actual location $(X, Y)^\top$ is given as:

$$\begin{pmatrix} X \\ Y \end{pmatrix} = Z \begin{pmatrix} \tan \theta \cos \phi \\ \tan \theta \sin \phi \end{pmatrix}, \quad (7)$$

where $\phi = \arctan(y, x)$ refers to the polar angle, which is shared by both p and P .

It can be seen that precise determination of the person location point p on the image plane is vital for estimating the corresponding physical position P . However, most previous approaches, restricted to axis-aligned [79, 60] or person-aligned [24] human-box representation, use the center of the detection box to approximate p . Few excep-

Method	val								test								FPS \uparrow
	mAP \uparrow	AP $_{50}$ \uparrow	AP $_{75}$ \uparrow	AP $_n$ \uparrow	AP $_m$ \uparrow	AP $_f$ \uparrow	AP $_{seen}$ \uparrow	AP $_{unseen}$ \uparrow	mAP \uparrow	AP $_{50}$ \uparrow	AP $_{75}$ \uparrow	AP $_n$ \uparrow	AP $_m$ \uparrow	AP $_f$ \uparrow	AP $_{seen}$ \uparrow	AP $_{unseen}$ \uparrow	
Seide <i>et al.</i> [60]	21.8	59.8	7.6	32.8	28.5	2.3	23.2	19.5	20.2	58.2	7.1	32.2	28.3	3.9	22.4	18.9	10.2
Li <i>et al.</i> [41]	28.5	63.3	20.1	46.8	24.2	1.3	33.8	29.3	27.2	65.2	21.3	47.2	24.8	1.3	31.8	27.6	0.6
Tamura <i>et al.</i> [64]	34.8	72.1	27.7	51.7	38.8	8.7	38.8	32.9	34.2	72.8	28.7	53.7	37.3	6.5	39.5	33.2	10.2
RAPiD [24]	40.3	77.9	34.8	55.3	41.9	9.2	44.7	37.6	39.2	77.9	35.4	54.8	40.1	7.9	44.2	37.3	8.4
Ours	47.2	82.3	48.2	63.8	54.1	14.0	50.6	45.5	46.2	81.1	47.3	66.1	53.5	12.6	49.3	44.9	12.1

Table 2: **Person detection results** on val and test sets of LOAF (§5.2).

Method	mAP \uparrow	AP $_{50}$ \uparrow	AP $_{75}$ \uparrow	Precision \uparrow	Recall \uparrow	F-Score \uparrow
Seide <i>et al.</i> [60]	20.9	50.6	10.2	80.6	39.5	53.0
Li <i>et al.</i> [41]	34.2	75.7	28.6	86.3	65.4	74.4
Tamura <i>et al.</i> [64]	29.3	61.0	23.4	88.8	51.2	65.0
RAPiD [24]	39.3	85.4	26.0	89.2	78.7	83.6
Ours	46.8	88.1	36.8	90.2	87.4	88.6

Table 3: **Person detection results** on CEPDOF [24] (§5.2).

Method	mAP \uparrow	AP $_{50}$ \uparrow	AP $_{75}$ \uparrow	Precision \uparrow	Recall \uparrow	F-Score \uparrow
Seide <i>et al.</i> [60]	16.1	39.4	9.0	70.9	38.6	50.0
Li <i>et al.</i> [41]	25.2	69.9	30.2	81.4	64.5	72.0
Tamura <i>et al.</i> [64]	28.8	59.8	24.2	77.0	52.4	62.4
RAPiD [24]	37.7	72.0	26.8	73.3	67.8	70.4
Ours	45.4	85.1	36.2	84.7	74.4	79.5

Table 4: **Person detection results** on WEPDToF [65] (§5.2).

tions [21], built on point-based human representation, treat the center of the detected human head as p . However, they suffer from a similar issue as the human head center is even often far from the exact position p human stand on, w.r.t. the image plane. Instead, with our radius-aligned human-box representation, the closest point to the principal point on the bounding box better corresponds to p .

4.3. Implementation Details

Network Architecture. Our fisheye person detector is built upon DAB-DETR [44], a prevalent variant of DETR [9] but converges much faster. Swin-T [45] is utilized as the backbone. For our LOAF with radius-aligned person boxes, the output of our detector is a 4D vector $\hat{b} \in [0, 1]^4$ that parameterizes 2D center coordinates, height, and width. Note that, for traditional fisheye detection datasets like [24, 65] with arbitrary-oriented person box annotations, an extra output dimension is needed for rotation angle regression.

Training Objective. Our fisheye person detector is end-to-end trained by jointly optimizing the vanilla detection loss used in DETR [9, 44] (referred as \mathcal{L}_{Det}) and our proposed rotation-equivariant constraint (*i.e.*, $\mathcal{L}_{\text{rotat-equiv}}$ in Eq. 5):

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda \mathcal{L}_{\text{rotat-equiv}}, \quad (8)$$

where the coefficient λ is empirically set to 0.5.

Training. Our fisheye detector is trained with a batch size of 8 for 50 epochs, where the AdamW [46] optimizer is employed with base learning rate $2e-4$ and decayed by 0.1 at epoch 40. The remaining hyper-parameters are determined following [64, 24]. Specifically, we initialize backbones with ImageNet [22] pre-trained weights and adopt standard data augmentation techniques, *i.e.*, color jitter, horizontal flip, and random scaling, with a base training size of 608×608 . For CEPDOF [24] and WEPDToF [65], we use a batch size of 128 and pre-train the detector on COCO [43] for 50 epochs to prevent over-fitting, as in [24, 65]. For the computation of our rotation-equivariant loss $\mathcal{L}_{\text{rotat-equiv}}$, the training images are rotated by a degree randomly sampled from 0 to 360. Our detector is implemented in PyTorch and trained on eight

NVIDIA Tesla V100 GPUs with a 32GB memory per-card. **Inference.** Once trained, our fisheye detector can be directly applied for locating persons on the omnidirectional image plane. After that, the physical locations of the detected persons can be easily obtained through the numerical solution described in §4.2. For fisheye person detection, we follow prior work [60, 41, 64, 24] to use 1024×1024 input resolution without any test-time augmentation or post-processing. Testing is conducted on a single NVIDIA V100 GPU with 16 GB memory.

5. Experiment

5.1. Experimental Setup

Evaluation Protocol. On the top of LOAF, we conduct experiments for fisheye based person detection and localization. Following the dataset splitting (*cf.* §3.3), performance are reported on *seen* and *unseen* scenes, respectively, for both val and test sets. This allows us to assess the generalization ability over different surveillance scenarios. For comprehensive study, we further report person detection performance on two existing fisheye datasets, CEPDOF [24] and WEPDToF [65]. Note that localization cannot be test since [24, 65] only provide person bounding box annotations.

Evaluation Metrics. For fisheye person detection, we follow COCO [43] to report the mean average precision (mAP) for IoU $\in [0.5:0.05:0.95]$. We also employ AP $_{50}$ and AP $_{75}$ for further analysis. For fisheye person localization, we measure positional error (PE) in meters, *i.e.*, Euclidean distance of calculated position and ground-truth position, as the conventions in visual localization [42, 58, 73, 79]. For detailed evaluation, on our LOAF, we report performance w.r.t. horizontal distance between human and the fisheye lens, *i.e.*, *near* (0~10 m), *middle* (10~20 m), and *far* (larger than 20 m). Hence we have AP $_{\{n,m,f\}}$ and PE $_{\{n,m,f\}}$ accordingly.

5.2. Performance on Person Detection

LOAF. We first report the person detection performance on LOAF. Specifically, four recent deep learning based fisheye

Method	val						test					
	mPE↓	PE _n ↓	PE _m ↓	PE _f ↓	PE _{seen} ↓	PE _{unseen} ↓	mPE↓	PE _n ↓	PE _m ↓	PE _f ↓	PE _{seen} ↓	PE _{unseen} ↓
Seide <i>et al.</i> [60]	1.298	0.561	1.332	3.109	1.206	1.382	1.321	0.706	1.309	3.482	1.306	1.386
Li <i>et al.</i> [41]	0.898	0.502	0.871	2.650	0.832	0.962	0.913	0.543	0.884	2.780	0.904	0.998
Tamura <i>et al.</i> [64]	0.755	0.429	0.736	1.862	0.709	0.821	0.778	0.471	0.826	2.160	0.724	0.836
RAPiD [24]	0.674	0.426	0.623	1.403	0.625	0.757	0.682	0.461	0.664	1.445	0.638	0.776
Ours	0.387	0.164	0.382	0.786	0.332	0.419	0.392	0.171	0.391	0.825	0.343	0.413

Table 5: **Person localization results** on val and test sets of LOAF (§5.3).

person detectors [60, 41, 64, 24] are involved for comparison. Among them, [60, 41] are training-free, thus their scores are obtained by directly running the algorithms on the val and test data. As [64, 24] are trainable methods, we first train them on the train set of LOAF following their official setups, and then report the scores on the val and test sets. Quantitative comparison results are presented in Table 2. Some essential observations are as follows:

- Our fisheye person detector significantly outperforms existing methods across all the metrics. For example, our detector provides a considerable performance gain in mAP, *i.e.*, **6.9%** and **7.0%** higher than the second best, RAPiD [24], on the val and test sets, respectively.
- Our detector not only handles nearby persons but also approaches distant targets well. Specifically, for the targets at 10~20 m away from the fisheye lens, our performance gain over other methods is more significant than the improvement reported with the persons at 0~10 m distance.
- Our detector also yields small performance gap between the seen and unseen scenes (*e.g.*, 49.3 → 44.9 in terms of mAP), validating our good generalizability.
- For those persons at very far distance (>20 m), our algorithm, though still giving higher scores than other competitors, suffers from great performance degradation. This sheds light on the direction of our future efforts.

CEPDOF [24]. CEPDOF contains eight videos with 25,504 frames in total. Following the official setup, we train our fisheye person detector on HABBOF [41] and MW-R [24] datasets, and report performance on CEPDOF. As shown in Table 3, our fisheye detector delivers state-of-the-art performance: it significantly outperforms RAPiD[24], the current top-leading algorithm, by **7.5%** in terms of mAP.

WEPDToF [65]. WEPDToF has 16 videos with 10,544 frames in total. Following the official setup, we use HABBOF [41], MW-R [24], and CEPDOF [24] for training and WEPDToF for testing. Table 4 summarizes the results. Impressively, our detector greatly suppresses all the other competitors across all the evaluation metrics.

5.3. Performance on Person Localization

Then we study the person localization performance on LOAF. None of previous fisheye person detectors [60, 41, 64, 24] are aware of the task of person localization. We therefore follow the common practice [79, 21] in the field of visual localization: for [60, 41, 64, 24], we project

Method	mAP↑	val seen			val unseen		
		mAP↑	AP ₅₀ ↑	AP ₇₅ ↑	mAP↑	AP ₅₀ ↑	AP ₇₅ ↑
Baseline [44]	43.1	46.4	81.2	48.4	40.9	76.2	38.9
+ Rotation Aug.	45.6	49.2	82.8	52.8	43.2	78.9	43.9
+ Rotation Equ.	47.2	50.6	83.7	54.6	45.5	80.4	47.1

Table 6: **Diagnostic results** on val set of LOAF (§5.4).

centers of their detected human boxes on the 3D world, based on the same strategy of ours (*cf.* §4.2), for physical position estimation. As shown in Table 5, our system produces much small localization errors in comparison with [60, 41, 64, 24], *i.e.*, **0.392 m** vs 0.682 m [24] and 0.773 m of [64] on the test set of LOAF. In addition, for all the methods, the localization performance is declined as target distance increases, but our system suffers from the smallest drop. Moreover, the results suggest there is still large room for improvement, thus we hope that our dataset could encourage continuous efforts in this challenging task.

5.4. Diagnostic Study

Table 6 studies the efficacy of our rotation equivariant training strategy (§4.1), on the val set of LOAF. Our baseline model [44] (*row* #1) gains 43.1% mAP. After adopting rotational data augmentation (*row* #2), the performance boosts by **2.5%** mAP. By contrast, our rotation equivariant training brings much larger improvements over the baseline, *e.g.*, **4.1%** mAP gain. It is remarkable that, training with standard rotational data augmentation technique can be viewed as a specific case of our equivariant training – only learning rotation-equivariant object-querying (*cf.* Eq. 4).

6. Conclusion

We presented LOAF – the first top-view fisheye dataset that supports large-scale study for person localization in realistic surveillance scenarios. With radius-aligned person-box annotations and precise location ground-truths, LOAF closes a crucial gap in the literature as these cases are not covered by previous datasets and annotation protocols. We further proposed an efficient fisheye person detection model that is equipped with a rotation-equivariant training strategy. The physical locations of detected persons are formulated based on the fisheye model and the altitude of the camera. We empirically verified the effectiveness and promising performance of our algorithm.

		# 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 [53] utilized in vanilla DETR [9] 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
λ: the balance factor
"""
def rotat_equi_training(I, gt):
    # F(I)
    m1 = Encoder(I)
    angle = randint(0, 360)
    # F(g^r(I))
    m2 = Encoder(rotate(I, angle))

    # {q_n}_{n=1}^N = κ(I)
    query1 = gen_proposal(m1)
    # {q_n^r}_{n=1}^N = κ(I^{g^r})
    query2 = gen_proposal(m2)

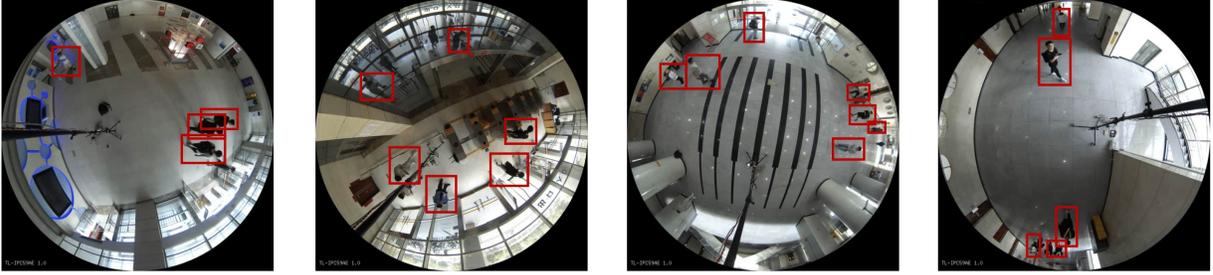
    # D(I, {q_n}_{n=1}^N)
    det1 = Decoder(m1, query1)
    # D(g^r(I), {q_n^r}_{n=1}^N)
    det2 = Decoder(rotate(m1, angle), query2)

    # L_det({b_{ℓ(n)}}_{n=1}^N, D(I, {q_n}_{n=1}^N))
    loss1 = det_loss(det1, gt)
    # L_det({b_{ℓ(n)}^r}_{n=1}^N, D(g^r(I), {q_n^r}_{n=1}^N))
    loss2 = det_loss(det2, label_rotate(gt, angle))

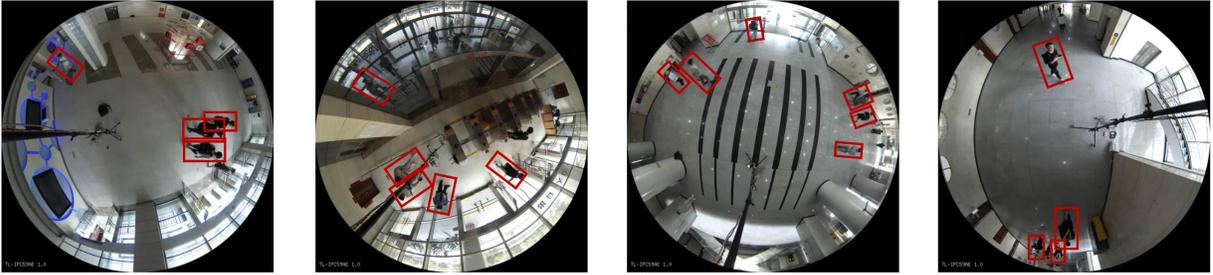
    return loss1 + λ*loss2

```

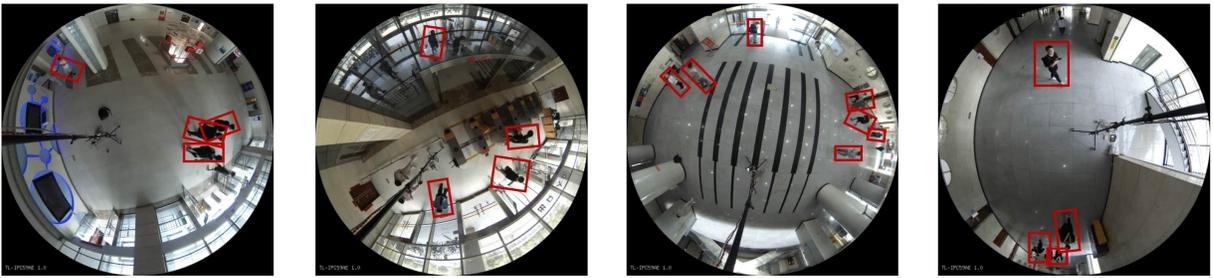
Seide et al. [60]



Li et al. [41]



Tamura et al. [64]



RAPiD [24]

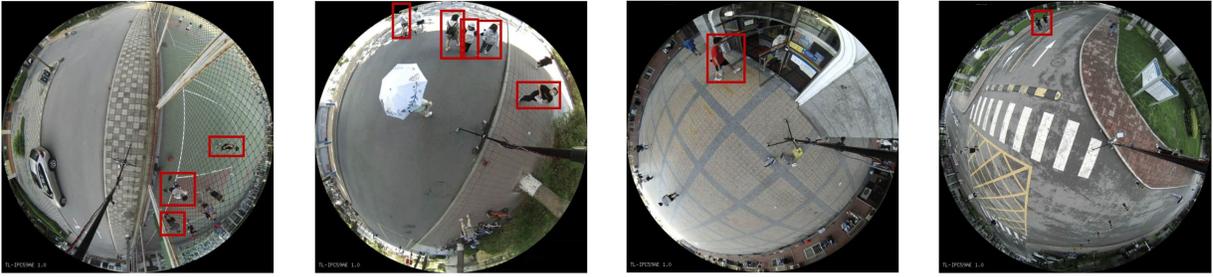


Ours

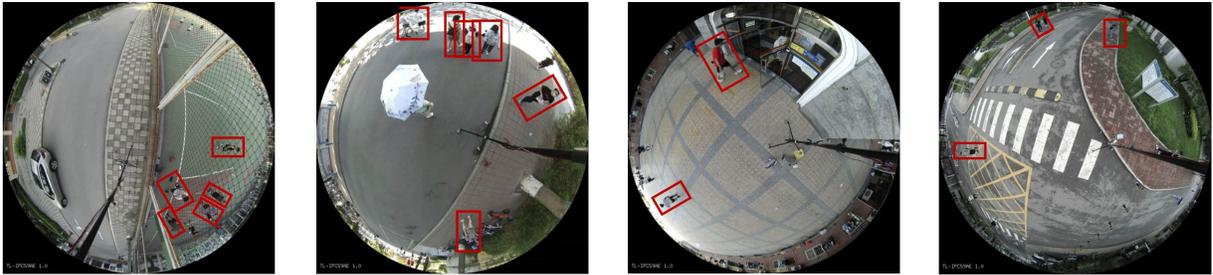


Figure S1: Visual comparison of detection results on the test set of LOAF. \diamond indicates targets missed by our method.

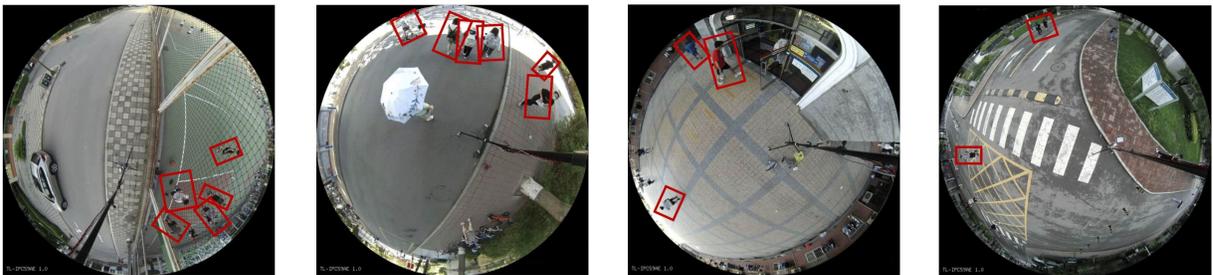
Seide et al. [60]



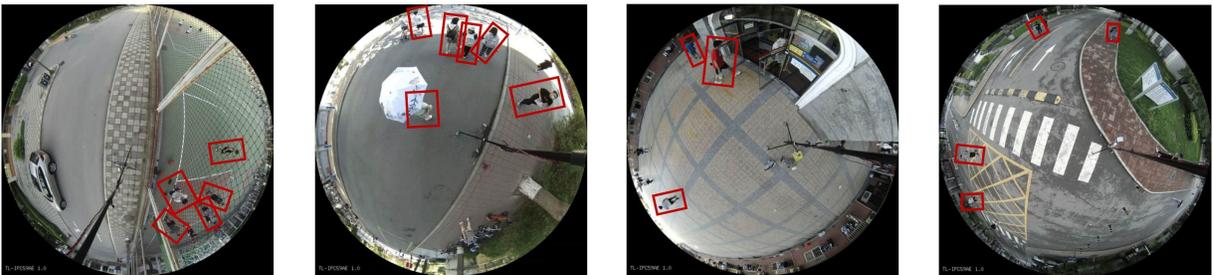
Li et al. [41]



Tamura et al. [64]



RAPiD [24]



Ours

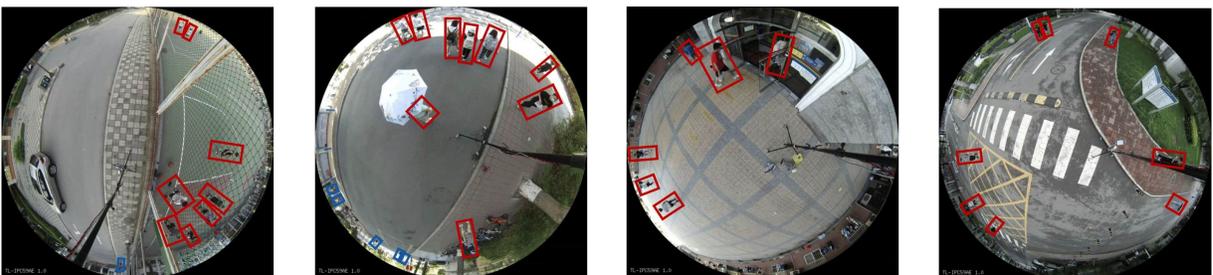
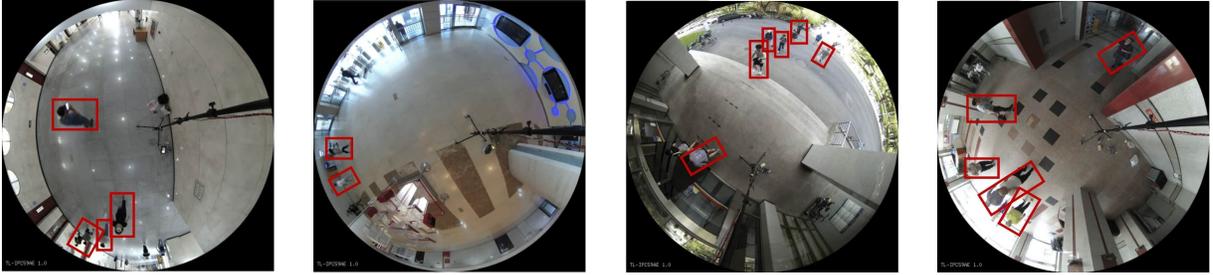


Figure S2: Visual comparison of detection results on the test set of LOAF. \diamond indicates targets missed by our method.

Seide et al. [60]



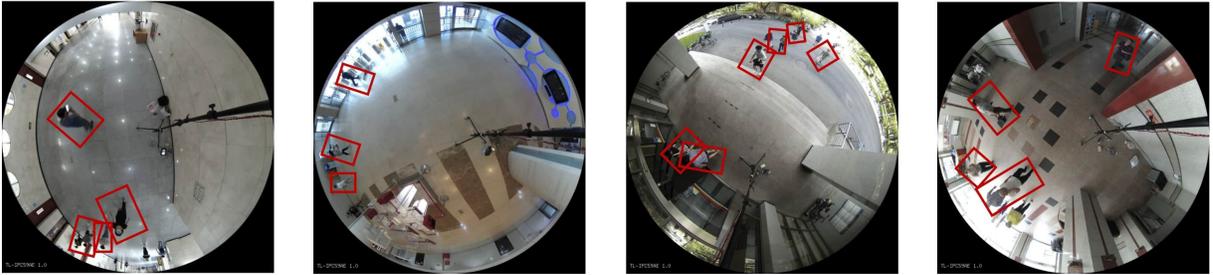
Li et al. [41]



Tamura et al. [64]



RAPiD [24]

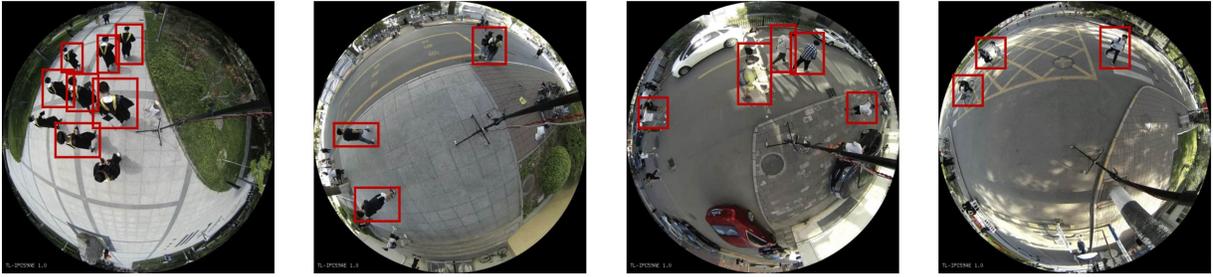


Ours

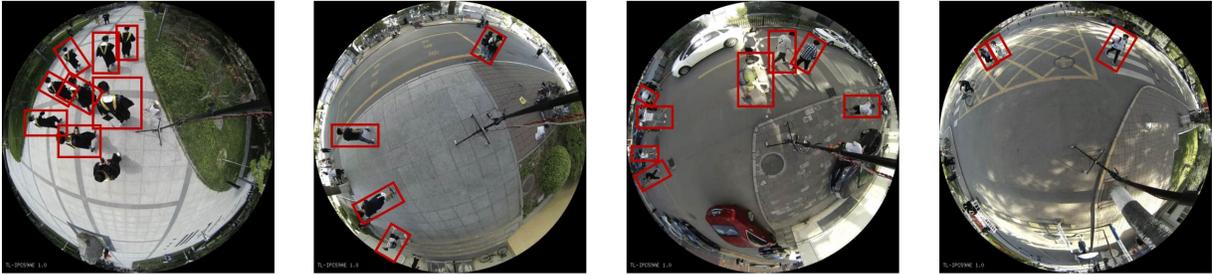


Figure S3: Visual comparison of detection results on the test set of LOAF. \diamond indicates targets missed by our method.

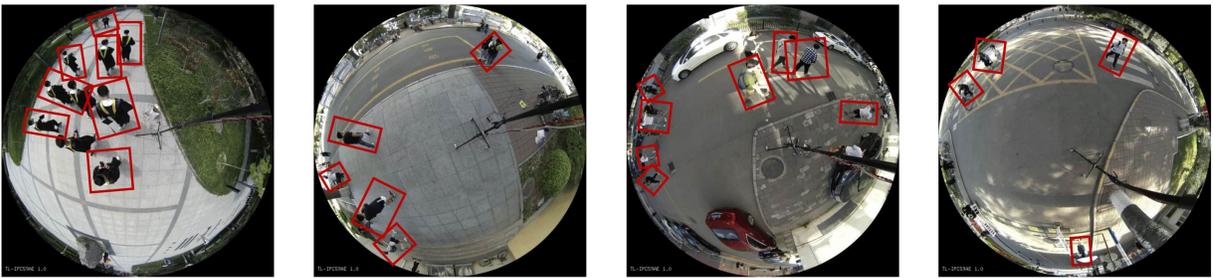
Seide et al. [60]



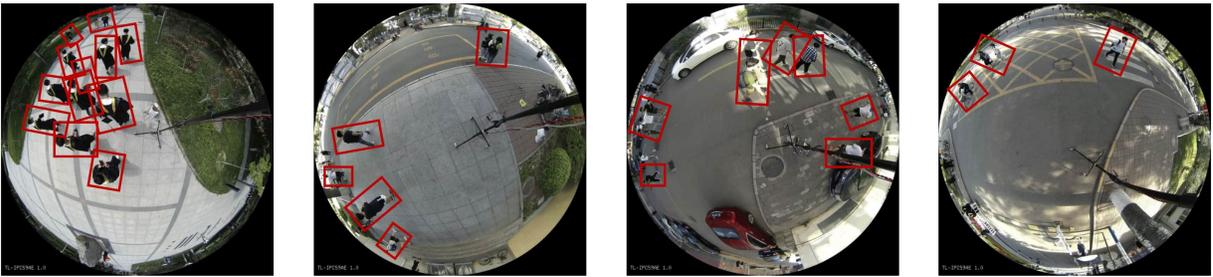
Li et al. [41]



Tamura et al. [64]



RAPiD [24]



Ours

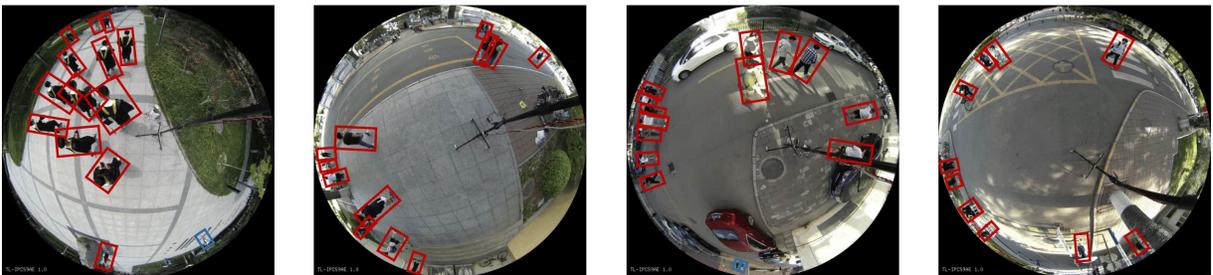


Figure S4: Visual comparison of detection results on the test set of LOAF. \diamond indicates targets missed by our method.

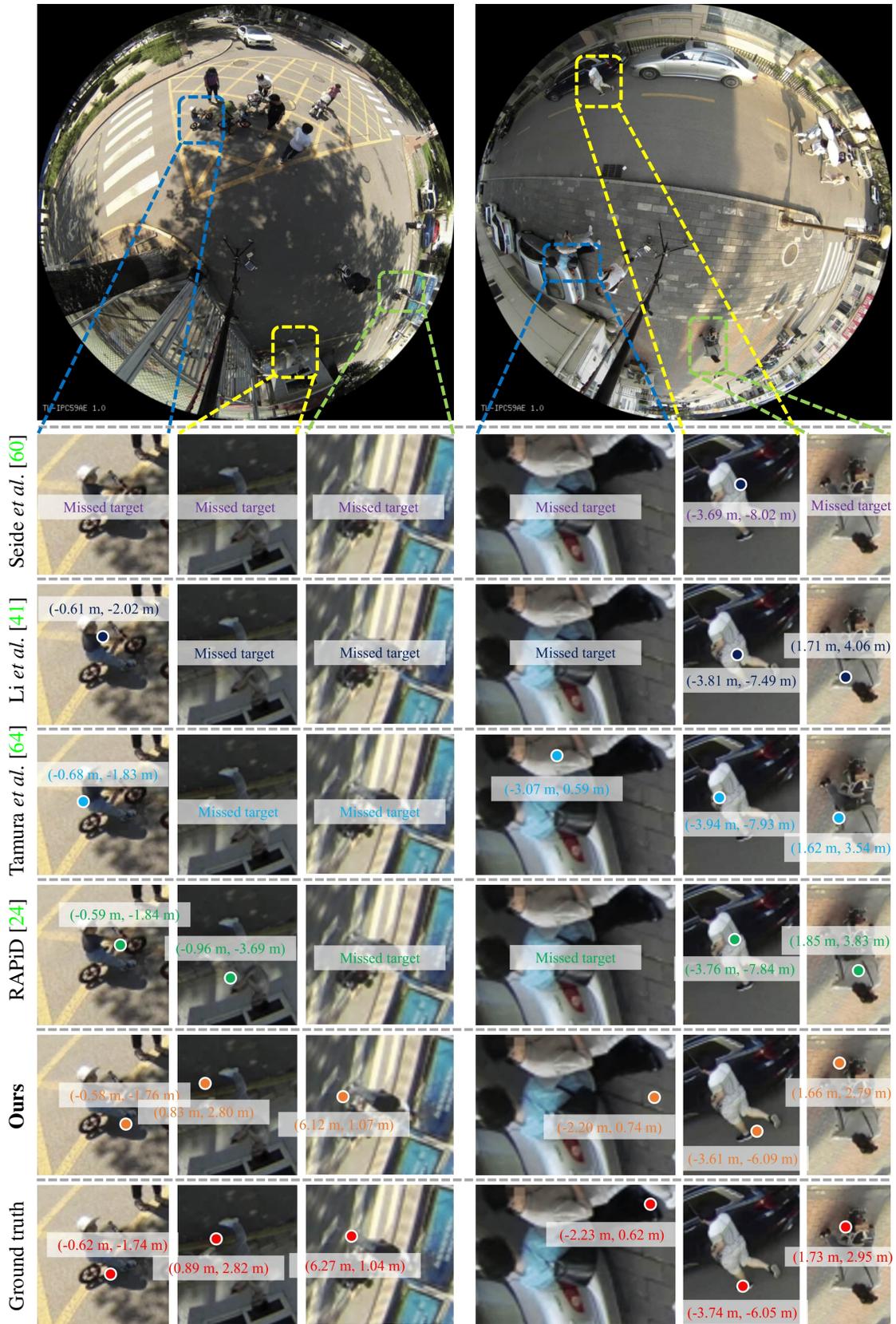


Figure S5: Visual comparison of localization results on the test set of LOAF. We selected three targets per frame for clear visualization.



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

References

- [1] Abdulrahman Alarifi, AbdulMalik Al-Salman, Mansour Al-saleh, Ahmad Alnafessah, Suheer Al-Hadhrami, Mai A Al-Ammar, and Hend S Al-Khalifa. Ultra wideband indoor positioning technologies: Analysis and recent advances. *Sensors*, 16(5):707, 2016. 2
- [2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 2016. 2
- [3] Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *ECCV*, 2018. 2
- [4] Federico Boniardi, Abhinav Valada, Rohit Mohan, Tim Caselitz, and Wolfram Burgard. Robot localization in floor plans using a room layout edge extraction network. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019. 2
- [5] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *CVPR*, 2017. 2
- [6] Eric Brachmann and Carsten Rother. Expert sample consensus applied to camera re-localization. In *ICCV*, 2019. 3
- [7] Guillaume Bresson, Li Yu, Cyril Joly, and Fabien Moutarde. Urban localization with street views using a convolutional neural network for end-to-end camera pose regression. In *IEEE Intelligent Vehicles Symposium*, 2019. 2
- [8] Jiale Cao, Yanwei Pang, Jin Xie, Fahad Shahbaz Khan, and Ling Shao. From handcrafted to deep features for pedestrian detection: a survey. *IEEE TPAMI*, 2021. 3
- [9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 5, 6, 7, 9
- [10] Bao Xin Chen and John K Tsotsos. Fast visual object tracking with rotated bounding boxes. *arXiv preprint arXiv:1907.03892*, 2019. 4
- [11] David M Chen, Georges Baatz, Kevin Köser, Sam S Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, et al. City-scale landmark identification on mobile devices. In *CVPR*, 2011. 3
- [12] Long Chen, Shaobo Lin, Xiankai Lu, Dongpu Cao, Hangbin Wu, Chi Guo, Chun Liu, and Fei-Yue Wang. Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey. *IEEE TITS*, 2021. 3
- [13] Liang Chen, Ling Pei, Heidi Kuusniemi, Yuwei Chen, Tuomo Kröger, and Ruizhi Chen. Bayesian fusion for indoor positioning using bluetooth fingerprints. *Wireless Personal Communications*, 2013. 1, 2
- [14] An-Ti Chiang and Yao Wang. Human detection in fish-eye images using hog-based detectors over rotated windows. In *ICME Workshop*, 2014. 3

- [15] Sheng-Ho Chiang, Tsaipei Wang, and Yi-Fu Chen. Efficient pedestrian detection in top-view fisheye images using compositions of perspective view patches. *Image and Vision Computing*, 2021. 3
- [16] Shin-Fang Ch'ng, Naoya Sogi, Pulak Purkait, Tat-Jun Chin, and Kazuhiro Fukui. Resolving marker pose ambiguity by robust rotation averaging with clique constraints. In *ICRA*, 2020. 2
- [17] Ibrahim Cinaroglu and Yalin Bastanlar. A direct approach for human detection with catadioptric omnidirectional cameras. In *Signal Processing and Communications Applications Conference*, 2014. 3
- [18] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICLR*, 2016. 5
- [19] Creative Commons. Attribution-noncommercial-sharealike 4.0 international. <https://creativecommons.org/licenses/by-nc-sa/4.0/>. 4
- [20] Joseph DeGol, Timothy Bretl, and Derek Hoiem. Chromatag: A colored marker and fast detection algorithm. In *ICCV*, 2017. 2
- [21] Carlos R del Blanco, Pablo Carballeira, Fernando Jau-reguizar, and Narciso García. Robust people indoor localization with omnidirectional cameras using a grid of spatial-aware classifiers. *Signal Processing: Image Communication*, 2021. 2, 3, 4, 5, 7, 8
- [22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7
- [23] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. Camnet: Coarse-to-fine retrieval for camera re-localization. In *ICCV*, 2019. 2
- [24] Zhihao Duan, Ozan Tezcan, Hayato Nakamura, Prakash Ishwar, and Janusz Konrad. Rapid: rotation-aware people detection in overhead fisheye images. In *CVPR Workshop*, 2020. 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14
- [25] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning SO(3) equivariant representations with spherical cnns. In *ECCV*, 2018. 5
- [26] Zahid Farid, Rosdiadee Nordin, and Mahamod Ismail. Recent advances in wireless indoor localization techniques and system. *Journal of Computer Networks and Communications*, 2013. 2
- [27] Mark Fiala. Artag, a fiducial marker system using digital techniques. In *CVPR*, 2005. 2
- [28] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3
- [29] Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. Neural reprojection error: Merging feature learning and camera pose estimation. In *CVPR*, 2021. 3
- [30] Yanying Gu, Anthony Lo, and Ignas Niemegeers. A survey of indoor positioning systems for wireless personal networks. *IEEE Communications surveys & tutorials*, 2009. 2
- [31] Janne Haverinen and Anssi Kemppainen. Global indoor self-localization based on the ambient magnetic field. *Robotics and Autonomous Systems*, 2009. 2
- [32] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *CVPR*, 2008. 2
- [33] Joel A Hesch and Stergios I Roumeliotis. A direct least-squares (dls) method for pnp. In *ICCV*, 2011. 3
- [34] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *ICANN*, 2011. 5
- [35] Danying Hu, Daniel DeTone, and Tomasz Malisiewicz. Deep charuco: Dark charuco marker pose estimation. In *CVPR*, 2019. 2
- [36] Juho Kannala and Sami S Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE TPAMI*, 28(8):1335–1340, 2006. 6
- [37] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *CVPR*, 2017. 2, 3
- [38] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 1, 2, 3
- [39] Oded Krams and Nahum Kiryati. People detection in top-view fisheye imaging. In *IEEE International Conference on Advanced Video and Signal based Surveillance*, 2017. 3
- [40] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *CVPR*, 2015. 5
- [41] Shengye Li, M Ozan Tezcan, Prakash Ishwar, and Janusz Konrad. Supervised people counting using an overhead fish-eye camera. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2019. 2, 3, 4, 5, 7, 8, 10, 11, 12, 13, 14
- [42] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3d point clouds. In *ECCV*, 2012. 1, 3, 7
- [43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 7
- [44] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *ICLR*, 2022. 7, 8
- [45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 7
- [46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 7
- [47] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 3
- [48] Kenro Miyamoto. Fish eye lens. *JOSA*, 54(8):1060–1061, 1964. 6
- [49] Jun Qi and Guo-Ping Liu. A robust high-accuracy ultrasound indoor positioning system based on a wireless sensor network. *Sensors*, 17(11):2554, 2017. 2
- [50] Hazem Rashed, Eslam Mohamed, Ganesh Sistu, Varun Ravi Kumar, Ciaran Eising, Ahmad El-Sallab, and Senthil Yogamani. Generalized object detection on fisheye cameras for autonomous driving: Dataset, representations and baseline. In *WACV*, 2021. 2
- [51] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster,

- stronger. In *CVPR*, 2017. 3
- [52] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3
- [53] Hamid Rezaatoughi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union. 2019. 9
- [54] Samer S Saab and Zahi S Nakad. A standalone rfid indoor positioning system using passive tags. *IEEE Transactions on Industrial Electronics*, 2010. 2
- [55] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 3
- [56] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *CVPR*, 2021. 2
- [57] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *CVPR*, 2018. 2
- [58] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *CVPR*, 2017. 2, 3, 7
- [59] Gerald Schweighofer and Axel Pinz. Robust pose estimation from a planar target. *IEEE TPAMI*, 28(12):2024–2030, 2006. 2
- [60] 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. 2, 3, 6, 7, 8, 10, 11, 12, 13, 14
- [61] Steve Shafer, John Krumm, Barry Brumitt, Brian Meyers, Mary Czerwinski, and Daniel Robbins. The new easyliving project at microsoft research. In *Proceedings of the 1998 DARPA/NIST Smart Spaces Workshop*, 1998. 2
- [62] Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A dataset for benchmarking image-based localization. In *CVPR*, 2017. 1
- [63] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, 2018. 1
- [64] Masato Tamura, Shota Horiguchi, and Tomokazu Murakami. Omnidirectional pedestrian detection by rotation invariant training. In *WACV*, 2019. 2, 3, 7, 8, 10, 11, 12, 13, 14
- [65] Ozan Tezcan, Zhihao Duan, Mertcan Cokbas, Prakash Ishwar, and Janusz Konrad. Wepdtof: A dataset and benchmark algorithms for in-the-wild people detection and tracking from overhead fisheye cameras. In *WACV*, 2022. 2, 4, 7, 8
- [66] Gonzalo Vaca-Castano, Amir Roshan Zamir, and Mubarak Shah. City scale geo-spatial trajectory estimation of a moving camera. In *CVPR*, 2012. 2
- [67] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. In *ICCV*, 2017. 2, 3
- [68] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019. 4
- [69] Tsaipei Wang, Chia-Wei Chang, and Yu-Shan Wu. Template-based people detection using a single downward-viewing fisheye camera. In *International Symposium on Intelligent Signal Processing and Communication Systems*, 2017. 3
- [70] Wenguan Wang, James Chenhao Liang, and Dongfang Liu. Learning equivariant segmentation with instance-unique querying. In *NeurIPS*, 2013. 5
- [71] Roy Want, Andy Hopper, Veronica Falcao, and Jonathan Gibbons. The active badge location system. *ACM Transactions on Information Systems*, 10(1):91–102, 1992. 2
- [72] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *ECCV*, 2016. 2
- [73] Aoran Xiao, Ruizhi Chen, Deren Li, Yujin Chen, and Dewen Wu. An indoor positioning system based on static objects in large indoor scenes by using smartphone cameras. *Sensors*, 2018. 1, 7
- [74] Chouchang Yang and Huai-Rong Shao. Wifi-based indoor positioning. *IEEE Communications Magazine*, 2015. 1, 2
- [75] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. Sanet: Scene agnostic network for camera localization. In *ICCV*, 2019. 2
- [76] Khalid Yousif, Yuichi Taguchi, and Srikumar Ramalingam. Monorgbd-slam: Simultaneous localization and mapping using both monocular and rgb-d cameras. In *ICRA*, 2017. 2
- [77] Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on google maps street view. In *ECCV*, 2010. 2
- [78] Amir Roshan Zamir and Mubarak Shah. Image geolocation based on multiplenearest neighbor feature matching using generalized graphs. *IEEE TPAMI*, 36(8):1546–1558, 2014. 2
- [79] Jun Zhu, Jiangcheng Zhu, Xudong Wan, Chao Wu, and Chao Xu. Object detection and localization in 3d environment by fusing raw fisheye image and attitude data. *Journal of Visual Communication and Image Representation*, 59:128–139, 2019. 2, 6, 7, 8