Applications of Deep Learning for Top-View Omnidirectional Imaging: A Survey

Jingrui Yu Ana Cecilia Perez Grassi Gangolf Hirtz
Chemnitz University of Technology, Germany
{jingrui.yu, ana-cecilia.perez-grassi, g.hirtz}@etit.tu-chemnitz.de

Abstract

A large field-of-view fisheye camera allows for capturing a large area with minimal numbers of cameras when they are mounted on a high position facing downwards. This top-view omnidirectional setup greatly reduces the work and cost for deployment compared to traditional solutions with multiple perspective cameras. In recent years, deep learning has been widely employed for vision related tasks, including for such omnidirectional settings. In this survey, we look at the application of deep learning in combination with omnidirectional top-view cameras, including the available datasets, human and object detection, human pose estimation, activity recognition and other miscellaneous applications.

1. Introduction

Omnidirectional cameras have the advantage of being able to capture a wide field of view (FOV). However, their projection models introduce a large distortion into their images. For this reason, computer vision methods developed for perspective images are not suitable for omnidirectional ones. In the last decade, computer vision has experienced a great advance thanks to the development of deep neural networks and the availability of large databases. However, this advance has focused almost exclusively on perspective images, both in the development of architectures and in the collection and annotation of data. It has not been until recent years that deep learning has begun to reach omnidirectional image processing, by collecting datasets and adapting existing architectures or developing new ones for this type of image.

Omnidirectionality can be achieved by using catadioptric, dioptric or polydioptric cameras. Catadioptric cameras combine a normal camera with a shaped mirror [35,66,101]. This mirror provides omnidirectionality as a surround-view, but the camera itself occludes the central part of the image. This problem is solved by dioptric cameras, which use a fisheye lens instead of a mirror. Finally, polydioptric cameras capture a spherical field of view by combining multiple cameras in a setup [8, 54].

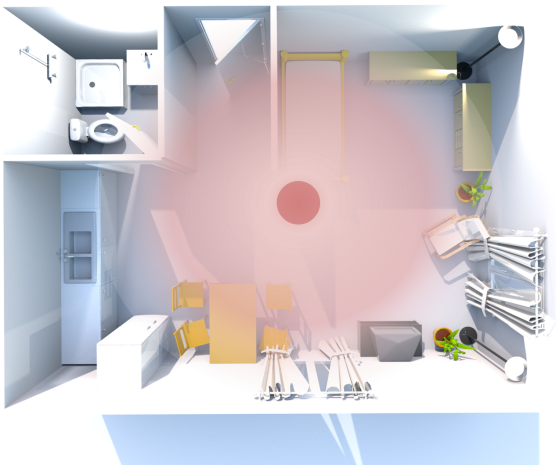
Especially dioptric cameras are gaining attention in many applications because they are simple and inexpensive. Depending on the task, these cameras can be mounted with a frontal view, as for example in driving applications [51, 72, 104, 106], with a vertical view as in teleconference applications [70] or with a top view as in surveillance applications [52, 56, 64, 67]. Also, their use for 3D-reconstruction, using one or more cameras, increases in the recent years [55, 82, 102]. In this survey we focus on deep learning algorithms developed for fisheye images captured from a top view. This kind of images are essential in surveillance and Ambient Assisted Living (AAL) applications [84], where the main research areas include person and object detection and human pose estimation.

Although there exist other surveys that focus on omnidirectional fisheye images, such as [2, 19, 50], they mostly discuss the frontal view. Other surveys of top-view imaging [1, 58] concentrate only on one application of the top-view perspective and do not specify the usage of omnidirectional cameras. The methods surveyed are mostly confined to classical computer vision algorithms. Therefore, we consider our survey essential for grasping the trend in applications of the combination of top-view fisheye imaging and deep learning.

This paper is organized as follows: in Sec. 2 we describe the camera geometry and the top view setup in detail. The available omnidirectional datasets are presented in Sec. 3. Sections 4, 5 and 6 cover object detection, pose estimation, human activity recognition and other miscellaneous applications. We conclude the survey in Sec. 7.

2. The omnidirectional top-view setup

As introduced in Sec. 1, this survey focuses of the top-view omnidirectional vision utilizing one or multiple dioptric cameras. The camera or the camera rig consisting of multiple cameras is usually hung on the ceiling near the center of the room. Figure 1 shows an example of such setting. Do not confuse this with the top view or bird's-eye-view in autonomous driving applications [85], which is synthesized from the surround view images.



(a) The top-view omnidirectional set-up in a one-room apartment. The red dot in the center illustrates the position of the camera.

(b) Example output of the set-up in a synthetic environment.

Figure 1. The top-view omnidirectional set-up and its output.

To utilize the advantages and tackle the shortcomings of this set-up, we need to understand the model of fisheye camera. An ideal fisheye lens can be described with the equidistant projection in Eq. (1). θ is the angle between the principal axis and the incoming ray, r is the distance between the image point and the principal point, and f is the focal length (see Fig. 2).

$$r = f\theta \tag{1}$$

There are other projection models that are less frequently used, see Tab. 1.

A realistic fisheye camera is not perfect and require a more complex model for calibration and precise image unwrapping and odometry. The publication [47] describes a generic model that enables the calibration of a fisheye camera with a single planar calibration pattern. DeepCalib presents the possibility of using DL for acquiring calibration parameters of fisheye cameras [7].

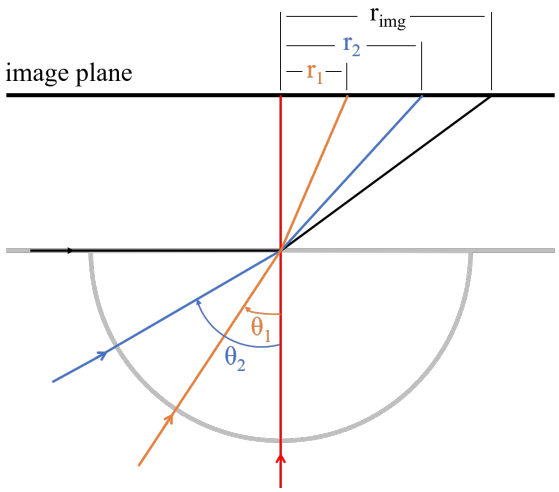


Figure 2. The equidistant projection. The red ray is the principle axis. The black ray is from the maximum visible angle and r_{img} is the radius of the image area.

Table 1. Other projection models used for fisheye cameras besides equidistant projection.

Projection type	Math. expression
Equisolid	$r = 2f \sin(\theta/2)$
Stereographic	$r = 2f \tan(\theta/2)$
Orthographic	$r = f \sin\theta$

An omnidirectional camera can be built with a normal CCD or CMOS camera and a fisheye lens [76]. Besides, there are commercial products from various companies, such as the Quasar™ Hemispheric Mini-Dome by TELEDYNE FLIR, the HemiStereo™ DK1 and NX by 3DVisionlabs, the panoramic series by HIKVISION, the IP Fisheye series by ABUS, the C71 and Q71 by MOBOTIX, the FE series by VIVOTEK, etc. They can be most easily found under the term "hemispherical camera" with a search engine.

3. Datasets

In deep learning approaches, the availability of quality data and ground truth annotations is essential for the training process. Today, models of almost every architecture can be obtained with weights trained on popular large-scale image datasets such as MS COCO [60], ImageNet [79, 105], etc. This enables users to avoid long and costly trainings from scratch and to facilitate the transfer learning ability of neural networks to adapt these exist-

ing models to new domains and new tasks. However, the images in these datasets are mostly collected with a perspective camera from a frontal view. Therefore, the pictured objects present a different appearance from those in top-view omnidirectional images, especially when they are close to the camera. This prevents not only the direct use of these models in such images, but also makes transfer learning extremely difficult, if a large amount of omnidirectional data for fine-tuning is not available. In the early 2010s, omnidirectional image datasets were scarce and insufficient for training or fine-tuning complex architectures. However, with the increasing interest in using omnidirectional cameras, the first real-world and synthetic datasets of top view fisheye images were created. Unlike general-purpose image datasets for classification, object recognition or segmentation, which are collected on the internet and have great intra-dataset variability, omnidirectional datasets are mostly recorded in a specific setting for a specific task. Therefore, they are usually continuous sequences extracted from videos, and the variability between images is lower. In the following subsections, we present and describe these datasets. Tab. 2 summarizes their technical characteristics. Links to the datasets in this chapter are accessed on March 17th, 2023.

3.1. Real-world datasets

The **Bomni Database**[1] (Boğaziçi University Multi-Omnidirectional Video Tracking Database) [24] is one of the earliest datasets of omnidirectional fisheye camera images. Although the authors list a few other datasets, they are not available anymore at the time of this review. Bomni DB is recorded for the purpose of human tracking in indoor scenes. Two fisheye cameras, one mounted on the ceiling and the other on a side wall, are used to simultaneously record two scenarios with a resolution of 640×480 pixels and a frame rate of 8 fps. Scenario 1 shows a single subject entering a room and performing six different actions before leaving. For this scenario a total of 10 videos with 5 different subjects are recorded. Scenario 2 presents 36 videos of multiple persons interacting in the same room. For this scenario a total of five actions are defined. The dataset provides tracking IDs, bounding boxes for moving subjects and action labels as annotations, which are given in vatic [94] format. It is to be noted that a portion of the annotations are generated from automatic tracking and interpolation. This often results in slight misalignment between the subject and its bounding box. Additionally, Bomni DB lacks labels for quasi-static persons in the scene.

HDA Person Dataset[2] [30] is a dataset for surveillance. Most of the image data are recorded by classic surveillance cameras, but the sequence *Cam 02* is recorded by a fisheye camera mounted on the ceiling of an elevator waiting area. The sequence has a resolution of 640×480 and a frame rate of 5 fps. In total 9819 frames are recorded. Bounding boxes of persons are provided in this dataset. Heavy motion blur is present throughout this recording.

PIROPO database[3] (People in Indoor ROoms with Perspective and Omnidirectional cameras) [21] is recorded simultaneously using a ceiling-mounted fisheye camera and a normal perspective camera. The scenes consist of a single person or multiple people walking, standing or sitting in a room. There are no interactions between the persons. It is a large scale dataset with over 100,000 annotated frames and a number of unannotated frames. The annotation is provided in the form of points, which mark the head positions of the persons in the image. In the work [107], the authors mention they down-sampled the original dataset with annotations and manually annotated the resulting dataset with bounding boxes for the persons.

MW-18Mar Dataset[4] from Mirror Worlds Challenge is an indoor top-view fisheye video dataset which consists of 30 videos an 13k frames. The original dataset are annotated with axis-aligned bounding boxes. For tracking purpose there are also annotated track trajectories. There are 3 main scenarios in this dataset: an observation room, a hallway and a synthetic scene of an observation room. The train set of this dataset is later annotated with rotated bounding boxes by the authors of [27] and is named **MW-R**[5].

Tamura *et al.* annotated Bomni, PIROPO and MW-18Mar datasets with rotated bounding boxes for their work [90]. The annotation files are in Pascal VOC format [29] and available online[6].

HABBOF (Human-Aligned Bounding Boxes from Overhead Fisheye Cameras) [56], **CEPDOF** (Challenging Events for Person Detection from Overhead Fisheye Images) [27] and **WEBDTOF** (In-the-Wild Events for People Detection and Tracking from Overhead Fisheye Cameras) [91] are datasets collected by the Visual Information Processing Laboratory of Boston University[7]. HABBOF provides two indoor scenes of 5837 frames. The annotations are given as bounding boxes aligned to the human body, which appear mostly in line with the radial axis of the omnidirectional image. CEPDOF is an extension of HABBOF. It provides 8 video sequences of different lev-

[1] https://www.cmpe.boun.edu.tr/pilab/pilabfiles/databases/bomni/

[2] https://vislab.isr.tecnico.ulisboa.pt/hda-dataset/

[3] https://sites.google.com/site/piropodatabase/

[4] https://www2.icat.vt.edu/mirrorworlds/challenge/index.html

[5] https://vip.bu.edu/projects/vsns/cossy/datasets/mw-r/

[6] https://github.com/hitachi-rd-cv/omnidet-rotinv

[7] https://vip.bu.edu/projects/vsns/cossy/datasets

Table 2. Technical attributes of omnidirectional image and video datasets. Types include real-world (R), synthetic (S) or hybrid (R+S). Res. stands for resolution and the unit is megapixels (MP). Year indicates the year of publication, not the year of appearance.

Dataset	Type	# of frames	Res. (MP)	Annotations	Classes	Year
Bomni	R	10,340	0.3	bbox, tracking ID, actions	9 actions	2012
HDA (*Cam 02*)	R	1,388	0.3	bbox	person	2013
LMS	R+S	515	1.2	–	–	2016
PIROPO	R	111,283	0.48	head point, bbox (3rd party)	person	2021
MW-18Mar	R+S	14,040	1.1 to 2.2	bbox, rotated bbox (3rd party)	person	2018
HABBOF	R	5,837	4.2	rotated bbox	person	2019
CEPDOF	R	25,504	1.1 to 4.2	rotated bbox	person	2020
WEBDTOF	R	10,544	0.6 to 5	rotated bbox	person	2022
FRIDA	R	18,318	4.2	rotated bbox, person ID	person	2022
DEPOF	R	3,594	4.2	bbox, point location	person	2023
FES	R	301	2.8	bbox, instance mask	6 classes	2020
360Action	R	784 clips	8.3	actions per video clip	19 actions	2020
FRailTRI20_DOD	R	44,099	0.93	temporal anomalies	7 anomalies	2020
OSD	R+S	39,200	1.0	bbox	person	2021
THEODORE	S	100,000	1.0	bbox, segmentation & instance mask	14 classes	2020
THEOStereo	S	31,250 pairs	1.0	depth map	–	2021

els of crowdedness under different lighting conditions. Unlike any earlier datasets, which are recorded in controlled settings, WEBDTOF is recorded in real-life situations. 14 scenes are recorded with different cameras and lens to form 16 videos. Thus, it covers common difficulties presented in real life: occlusions, camouflage, cropping, tiny people, non-circular FOV and children. The same research group also presents two datasets for other applications: **DEPOF** (Distance Estimation between People from Overhead Fisheye cameras) [62] and **FRIDA** (Fisheye Re-Identification Dataset with Annotations) [17]. DEPOF provides 3,526 frames for calibration purpose and 68 frames for training and testing of a person distance measurement method. FRIDA is destined for person re-identification, but rotated bounding boxes are also available.

FES[8] (Fisheye Evaluation Suite) [81] is an indoor dataset. It differs greatly from the afore mentioned datasets by providing bounding boxes and instance segmentation masks for 6 classes: person, TV, table, armchair, chair and wheeled walker. The disadvantage of this dataset is its relative small size at only 301 frames.

360action[9] [53] is a dataset for action recognition in the form of video clips. In each clip, which has the length of 6 to 10 seconds at 30 fps, a number of subjects perform daily actions by themselves or with interactions. Action labels are given for each video, without specifying the subject.

FRailTRI20_DOD (French Rail Technological Research Institute Door Obstacle Detection 2020) [52] was created specifically for the surveillance of door areas of trains. The images were captured at 20 fps. Seven anomalies regarding train doors and passengers were performed by five actors. The authors defined a set of annotations regarding instance positions and displacement, door state, hazardous events, pedestrian actions and combine them into a temporal segmentation of an event. The dataset as well as a meticulous annotation guide can be acquired by contacting the first author.

3.2. Synthetic datasets

Apart from the subset in MW-18Mar dataset, multiple synthetic datasets of omnidirectional images have been created for a variety of purposes.

LMS Fisheye Dataset[10] [28] provides a variety of synthetic and real-world fisheye image video sequences, among which *HallwayC*, *LivingroomB*, *Room*, *HallwayB*, *LivingroomA* and *LivingroomC* are top view. These sequences do not have corresponding annotations.

OSD[11] (Omnidirectional Synthetic Datasets) [5] is a dataset for person recognition and surveillance. Therefore, it only provides annotations for persons in the form of segmentation masks and bounding boxes. The persons are small in size to simulate actual surveillance situations where a large area is monitored by one omnidirectional camera. Besides Omnidirectional images it also provides rectified images.

THEODORE[12] (synTHEtic tOp-view inDoOR scEnes

[8] https://www.tu-chemnitz.de/etit/dst/forschung/comp_vision/datasets/fes/

[9] https://github.com/ryukenzen/360action

[10] https://www.lms.tf.fau.eu/research/downloads/fisheye-data-set/

[11] https://datasets.vicomtech.org/v4-osd/OSD_download.zip

[12] https://www.tu-chemnitz.de/etit/dst/forschung/comp_vision/datasets/theodore/

dataset) [81] provides a large-scale diverse dataset with annotations for semantic segmentation, instance segmentation and bounding boxes for object detection. **THEOStereo**[13] [86] is derived from THEODORE and aims to aid depth estimation using top-view fisheye cameras. It provides image pairs from two virtual cameras and the corresponding depth maps as ground truth annotations. The baseline of the stereo cameras is $0.3\,\mathrm{m}$.

3.3. Other datasets

Other datasets such as **DPI-T**[14] (Depth-Based Person Identification from Top View) [38], **TVPR**[15] (Top View Person Re-Identification) dataset [57] does not use fisheye cameras, but the overhead viewpoint is nonetheless useful for training networks applied for omnidirectional images. These two datasets contain a depth channel in addition to RGB channels. **PanopTOP31K**[16] [34] is a semisynthetic RGB dataset from the top-view for view-invariant 3D human pose estimation. This is the first large-scale dataset that features top-view human keypoints. However, its low resolution at 256×256 pixels and heavy artifacts affect its usability negatively.

4. Person and object detection

4.1. Person detection

The datasets in Sec. 3 clearly show that the main application of omnidirectional cameras is person detection and tracking. Using deep learning for person detection in omnidirectional images, typically a CNN-based object detector, faces a few obstacles. Firstly, standing people appear in line with the radial axis of the image rather than upwards in images from side-mounted cameras. Secondly, the equidistant projection of fisheye cameras results in considerable deformation of objects. These two problems restricts the utilization of pre-trained models and reduces the effectiveness of transfer learning. Additionally, when the person stands directly under the camera, it has a unique appearance that is unseen in normal perspective images. We review the researches to see how these problems are progressively solved. An overview is provided in Tab. 3. Note that due to the lack of a common large scale dataset, researchers have used different datasets and metrics for the evaluation, therefore it is not possible to compare the performances to each other directly. Thus, the performances are not listed in the table.

Nguyen *et al.* [67] combined Adaptive Gaussian Mixture Model (AGMM)-based background subtraction and a simple CNN inspired by Tiny Yolo [73] to perform pedestrian detection. The network takes the foreground mask and single-channel grayscale images as input. Their evaluation shows an AP of 0.86 at their house dataset.

Seidel *et al.* proposed **OmniDetector** [83], which uses the camera calibration parameters to unwrap one omnidirectional image into 94 highly overlapping perspective images and then apply the pre-trained YOLOv2 [74] to detect persons. The bounding boxes are projected back into the omnidirectional image using a look-up table (LUT). Non-maximum suppression (NMS) is applied to the detections to eliminate overlapping bounding boxes and generate the final detection. They achieved an AP@0.5IoU of 0.646 on PIROPO when using soft-NMS with Gaussian smoothing. This method enables the use of CNN-based detectors without the need for collecting new data and training. However, its shortcomings are obvious. It has a large overhead, partly because of the transformations and partly because of the large amount of inferences for one image. It cannot detect persons directly under the camera, since the network has not seen such examples. Furthermore, it requires that the camera calibration parameters are known, which isn't feasible at all times. Curiously, **Chiang *et al.*** [15] used the same approach in 2021, with the only improvement of reducing the number of ROIs to eight.

Li *et al.* proposed in [56] to use a rotating rectangular focus windows to extract a part of the image, which will be rotated to maintain the upright direction of the person. The maximum number of focus windows is 24. Then the detection is performed with YOLOv3 [75] and consecutive NMS. The authors used background subtraction to identify regions of interest (ROI) where people are present and discard the focus windows without human activity, thus reducing the computational cost. This method is tested on HABBOF with the F-score of 0.88. The similarities with OmniDetector are the usage of multiple overlapping windows and that the detector does not need to be re-trained or fine-tuned. But this method reduces the computational cost and does not require the camera parameters to be known. However, it still does not address the person-under-the-camera problem, and it is by design not able to detect stationary persons.

Yu *et al.* took another approach and tried to achieve person detection directly in omnidirectional images with **OmniPD** [107]. They presented a training paradigm, by which omnidirectional images are combined with a dataset of normal perspective, in this case the PASCAL VOC dataset [29], to finetune a CNN-based object detector. Random horizontal and vertical flipping and random 90-degree rotation was used as data augmentation to compensate for rotation variance in omnidirectional images. Their best result was achieved with SSD [61] at AP@0.5IoU at 0.863, albeit on their own dataset.

Tamura *et al.* tried to achieve pedestrian detection in

[13] https://www.tu-chemnitz.de/etit/dst/forschung/comp_vision/datasets/theostereo/
[14] https://github.com/zhengkang86/ram_person_id
[15] https://vrai.dii.univpm.it/re-id-dataset
[16] https://github.com/mmlab-cv/PanopTOP

Table 3. Person detection in Omnidirectional images

Architecture	Main algorithm
Nguyen *et al*. [67]	AGMM-background subtraction + tiny YOLO
OmniDetector [83]	Unwapping + YOLOv2 + NMS
Li *et al*. [56]	Rotating window + background subtraction + YOLOv3
Tamura *et al*. [90]	Rotation-invariant training + YOLOv2 + BBR
OmniPD [107]	Hybrid training + rotation augmentation
RAPiD [27]	YOLOv3-based network + orientation prediction head + angle aware loss
ARPD [64]	CenterNet + orientation prediction head + rotation aware loss function
Haggui *et al*. [37]	RAPiD + color histograms for tracking
Wang *et al*. [98]	Dual Mask R-CNN + image region separation + scene specific training
GSAC-DNN [32]	2D grid of simple CNN-classifiers
Callemein *et al*. [9]	Low resolution image + temporal interlacing kernel in YOLOv2
OmniDRL [71]	Deep Q-Network + camera calibration
Wiedemer *et al*. [100]	Faster-RCNN + few-shot training

omnidirectional images by training YOLOv2 with randomly rotated perspective images from COCO [90]. DPI-T dataset was also used in the training. To overcome YOLOv2's problem of generating overlapping bounding boxes, they proposed bounding box regression (BBR) based on mean shift clustering of the center points of bounding boxes. A simple position-based bounding box angle determination was added to the refinement process. The authors manually annotated MW-18Mar, PIROPO, Bomni and CVRG for evaluation, as mentioned in Sec. 3.1.

Duan *et al*. proposed **RAPiD** [27], a new YOLO-inspired network architechture, which predicts the rotation angle besides the usual position and size of the bounding boxes. To train this network they added rotation-angle loss to the loss function of YOLOv3. There network is first pre-trained on COCO, then finetuned on two of the three datasets they annotated (MW-R, HABBOF, CEPDOF) and tested on the remaining one. They reached AP@0.50IoU of 0.967, 0.981 and 0.858 for MW-R, HABBOF and CEPDOF, respectively. Their team further improved the performance by extending RAPiD with temporal information [91]. Minh *et al*. used the same strategy to extend CenterNet to predict human aligned bounding boxes, and named their architecture **ARPD** [64]. Their experiments show that their method reaches similar AP as RAPiD while doubling the inference speed. **Haggui *et al*.** used RAPiD for initial detection and added tracking by using the color histograms [37].

The most recent development is [100]. **Wiedemar *et al*.** proposed a few-shot adversarial training scheme for Faster-RCNN [77] so that a pre-trained detection model can be adapted for person detection in top-view omnidirectional images with less than 100 annotated training samples. The techniques they used include loss coupling, global and instance level feature alignment. Their method can achieve higher accuracy when the number of annotated samples are smaller than 100. A key difference to previously mentioned methods is that this method is aimed at adapting an existing model with minimum amount of effort to a certain use case, instead of trying to create a model with maximum generalization power. Therefore, cross-dataset evaluation by the author shows that the model loses generalization power when the number of training examples exceed 50.

Besides the common methods for detection, researches have experimented with other ways of person detection with special features. **Wang *et al*.** [98] proposed to use Mask-RCNN [41]. The advantage is that segmentation mask does not have the problem of not aligning with the orientation of the person. They divided the images into a central region and a peripheral ring. The peripheral ring is further divided into three sectors, which are then warped into rectangles and stacked together to form one square image. The detection is performed using two separate detectors, one for the central region and the other for the outer region. Fuertes *et al*. proposed a Grid of Spatial-Aware Classifiers [21] based on Deep Neural Networks (**GSAC-DNN**) [32]. A feature map is generated by a ResNet-32 [42] backbone. It is fed to a 2D grid of simple classifiers consisting of a convolution layer and a linear layer. The location of the person is calculated based on the confidence scores of the classifiers. GSAC-DNN is end-to-end trainable, however it can only detect the general position of the person but not a bounding box. The work of **Callemein *et al*.** [9] is intended for occupancy detection in meeting rooms or for flex-desking, yet the detection results are still presented as bounding boxes. They use extremely low resolution images of 96×96 pixels to preserve privacy. To compensate for the information loss caused by the low resolution, they implemented a temporal interlacing kernel, which combines multiple consecutive frames into one high resolution feature map. Their network is able to run on embedded systems such as the Raspberry Pi 3B [31] at 0.77 fps. Pais *et al*. used a deep Q-Net (DQN)-based [65] network and the camera calibration parameters

to perform person detection and predict the 3D position of the person in the world coordinate [71]. The employment of reinforced learning is quite unusual. Their implementation is named **OmniDRL**.

4.2. Object detection

Object detection in omnidirectional images has not been widely researched. One reason is that it is not as useful as person detection. Another reason is the lack of data for training. **Scheck *et al.*** created the THEODORE dataset [81] to solve this issue. THEODORE contains five classes besides person: armchair, chair, table, TV and wheeled walker. They trained SSD [61], R-FCN [20] and Faster R-CNN [77] using this dataset and tested the trained networks on FES dataset. The mAP for all six classes reached 0.613 with Faster R-CNN. They used THEODORE to further train the anchorless CenterNet [26] for object detection [80]. In this work, they introduced unsupervised domain adaptation (UDA) to bridge the gap between synthetic image domain and real-world image domain, which is typically used in semantic segmentation. CenterNet was extended with two methods of UDA: entropy minimization (EM) [95] and maximum squares loss (MSL) [14]. The unlabeled target dataset used in [80] is CEPDOF. With the UDA-extended CenterNet they raised the mAP on the same FES dataset to 0.690 and doubled the inference speed at the same time. Another reason that object detection in omnidirectional view is underexplored is perhaps the very limited use cases. However, this could still be useful for accomplishing complex tasks with omnidirectional images, such as action recognition for smart monitoring systems, for example, in [84], a system is built for monitoring elderlies with dementia.

5. Human pose estimation and activity recognition

Human pose estimation (HPE) refers to the process of finding the joints of a person and connecting them into a skeleton. Pose estimation is the second most researched application of omnidirectional images. In this section we take a look at the approaches for 2D and 3D pose estimation. We also check out the researches for Human Activity Recognition (HAR), which often follows pose estimation.

5.1. Pose estimation with overhead fisheye camera

Georgakopoulos *et al.* [22, 36] employ a 3D human model to create a dataset of binary silhouettes, which are rendered through the calibration of a fisheye camera. The CNN is trained to differentiate between the pre-set postures, rather than estimate the joint positions. **Denecke and Jauch** [25] use the 3D point cloud calculated by the smart sensor and prior knowledge of the human body to estimate the joint positions. The results of this method are restricted by factors such as the mounting position of the camera and differences between each individual body. The inference speed is limited by the speed of the smart sensor. **Heindl *et al.*** [43] generated rectilinear views of the area that contains human in an omnidirectional image, thus the person appears upright in the virtual view. OpenPose [10] is then applied to this virtual image to perform pose estimation. They used a pair of calibrated fisheye cameras to get two skeletons, which then are combined into a 3D skeleton by using Direct Linear Transform (DLT) [40] in the rectilinear views.

Though not using omnidirectional images, the following two works perform HPE for the top-view. Haque *et al*. train CNN and LSTM [44] to achieve view-point invariant 3d pose estimation on a singular *depth image* [38, 39]. Garau *et al*. achieve viewpoint-invariant 3D HPE with a capsule auto-encoder named DECA [33] on depth and RGB images, namely ITOP [39] and PanopTOP31K datasets.

5.2. Egocentric 3D pose estimation

Egocentric pose estimation is a special case of using fisheye cameras in the top-view. The camera is not mounted over the person, instead, it is mounted with an apparatus on the head of the person with a small horizontal distance. Visible to the camera is the front or the frontal side of the body, as well as the peripheral of the person. **EgoCap** is a dual fisheye camera set-up mounted on a bike helmet with either a T-shaped or a Y-shaped wooden frame [78]. The cameras extrude about $25\,\mathrm{cm}$ to the front of the carrier. The authors created a dataset using a motion capture system to create the ground truth and projected the joint locations into the images from their set-up. With this dataset they finetuned ResNet101, which is pre-trained on MPII [4] and Leed Sports Extended Dataset [46], to generate heatmaps for 18 joints. The 3D skeleton is constructed in real-time from the 2D skeleton and a 3D body model, which must be adjusted for each user. **Mo^{2}Cap^2** [103] and **xR-EgoPose** [93]**/ SelfPose** [92] are similar implementations with a single fisheye camera. Both works developed their own synthetic datasets for training. Mo^{2}Cap^2 used one branch of CNN to generate heatmaps of joints for the whole body and another branch for the zoomed-in lower body. With the help of a CNN that estimates the distance between the joints and the camera, the joints are finally projected into the 3D coordinates. The calibration information of the cameras are essential for accurate 3D pose estimation. xR-Egopose used ResNet101 for joint heatmap generation. A lifting module takes the heatmaps as input and regresses the 3D pose from them as well as outputs the 2D heatmaps in high resolution. **Wang *et al.*** proposed in [97] a method for estimating not only the local pose, but also the global pose, which means the 3D joint positions in the world coordinate system are estimated. Their pipeline makes use of image

sequences instead of inferencing on a single frame. At the same time, they utilized motion prior, which is learned from AMASS dataset [63], to reduce temporal jitter and unrealistic motions from the estimated poses. Their set-up is similar to Mo^{2}Cap^2 and xR-EgoPose by mounting a single fisheye camera onto a helmet. In [96] they further proposed to additionally use an external camera for weakly supervised training. Their dataset for this task is named EgoPW. **EgoGlass** [108] is more extreme in terms of minimizing the apparatus size. They mounted two cameras to a normal eyeglass, each of which records a side of the body. The pose estimation is solved in the stitched-together image. We also notice **Cha *et al.*** [11, 12] used similar set-ups for their implementation, but not fisheye cameras.

5.3. Action recognition

Li *et al.* [53] proposed to perform action recognition in top-view fisheye camera images. They first use Mask-RCNN to find the spine lines of standing persons in the image. The cross point of the spine lines are deemed the optical center of the fisheye camera and the spherical image is dewarped into a panoramic image around it. Camera calibration information is not necessary in this process and the panoramic image is set to a pre-defined size. The authors use Mask-RCNN to perform person detection in the panoramic image and max pool the bounding boxes across 16 frames in each clip to form the ROIs. A 3D ResNet is used for action recognition through the 16 frames. A binary mask, which is generated from the ROIs, is multiplied with the feature maps from the 3D ResNet to reduce calculation cost. They used Multi-instance Multi-label Learning (MIML) to train a network for estimating scores for a series of actions. This work is further developed by **Stephen *et al.*** [88] by adding a second parallel pipeline for persons in the central area. Instead of using the panoramic view, this pipeline directly generates stacked feature maps for each person in the omnidirectional image, in which person detection is performed by RAPiD [27].

6. Other applications

Except for the intensively researched topics, multiple applications of omnidirectional images exist, mainly due to its wider FOV. Researchers have applied deep learning to these applications, however, DL does not stand in the focus in these applications. Much effort is given to solve the unique challenge of the equidistant projection of fisheye cameras, as well as other domain-specific problems.

Laurendin *et al*. proposed to use a top-view camera for **anomaly detection in train door area for autonomous trains** [52]. They created a dataset, in which the door area of a train is simulated, and systematically annotated it, see Sec. 3.1. They adapted the network in [68]. The results are inconclusive. Kim *et al*. proposed to use multiple top-view fisheye cameras for **parking lot surveillance** to determine vehicle positions [48]. The cameras were first calibrated using RANSAC to obtain their intrinsic and extrinsic parameters. SegNet [6] was used to generate segmentation masks for vehicle detections. They developed a method to estimate the actual size of the vehicle based on the calibration parameters of the camera and the generated segmentation mask. To get precise groundtruth data, the authors built a $1/18$-scale test bench using model cars and wooden frames. Their method was tested with an average distance error of $0.24\,\mathrm{m}$ (scaled to real-life) for vehicle position estimation and an average direction error of $4.8°$ for vehicle moving direction. Akai *et al*. used a fisheye camera for **grape bunch counting** [3]. In contrary to other examples reviewed in this paper, this work uses the bottom-up view instead of the top-down view. But this is the same approach in essence. The difference of viewpoint is just because the ROI is over the head instead of on the ground. Another important application is **indoor livestock monitoring**, such as body segmentation, identification, behavior recognition. Using top-view omnidirectional cameras in such farming areas, which are usually large and densely packed with animals, provides an unoccluded view with minimum number of cameras. Chen *et al*. [13] provide a comprehensive review of this application area, including the use of top-view omnidirectional cameras and deep learning. Li *et al*. performed **3D room reconstruction** using a single top-view fisheye camera [55]. They used RefineNet [59] for semantic segmentation of the room to aid structural line selection. The final result of their method is a cuboid representation of the room.

7. Conclusion

We can conclude from our survey that the main application areas of top-view omnidirectional imaging are surveillance and AAL. Researchers have created a considerable amount of data to facilitate the development of deep learning algorithms. With this, the implementations of deep learning algorithms show very promising results.

An obviously underexplored research area is human pose estimation. Though it has been greatly advanced for perspective images, the transfer to omnidirectional images is slow due to the high expense related to collecting human keypoints data with reliable ground truth [45]. The recent development in novel-view synthesis such as A-Nerf [89] and HumanNerf [99] could be the solution to this problem. More researches in this area will benefit further applications, such as fall detection, where only rule-based methods have been explored [23, 49, 69, 87]. Another possible research direction could be the usage of network architectures that are specifically developed for the geometry of fisheye images and related projections, such as spherical CNNs [16] and SphereNet [18].

References

[1] Misbah Ahmad, Imran Ahmed, Kaleem Ullah, Iqbal khan, Ayesha Khattak, and Awais Adnan. Person detection from overhead view: A survey. *International Journal of Advanced Computer Science and Applications*, 10(4), 2019. Copyright - © 2019. This work is licensed under https://creativecommons.org/licenses/by/4.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2022-11-29. 1

[2] Hao Ai, Zidong Cao, Jinjing Zhu, Haotian Bai, Yucheng Chen, and Lin Wang. Deep learning for omnidirectional vision: A survey and new perspectives, 2022. 1

[3] Ryota Akai, Yuzuko Utsumi, Yuka Miwa, Masakazu Iwamura, and Koichi Kise. Distortion-adaptive grape bunch counting for omnidirectional images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 599–606. IEEE, 2021. 8

[4] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 7

[5] Nerea Aranjuelo, Sara García, Estíbaliz Loyo, Luis Unzueta, and Oihana Otaegui. Key strategies for synthetic data generation for training intelligent systems based on people detection from omnidirectional cameras. *Computers & Electrical Engineering*, 92:107105, 2021. 4

[6] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 8

[7] Oleksandr Bogdan, Viktor Eckstein, Francois Rameau, and Jean-Charles Bazin. Deepcalib: a deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–10, 2018. 2

[8] Josep Bosch, Klemen Istenič, Nuno Gracias, Rafael Garcia, and Pere Ridao. Omnidirectional multicamera video stitching using depth maps. *IEEE Journal of Oceanic Engineering*, 45(4):1337–1352, 2020. 1

[9] Timothy Callemein, Kristof Van Beeck, and Toon Goedemé. Anyone here? smart embedded low-resolution omnidirectional video sensor to measure room occupancy. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1993–2000. IEEE, 2019. 6

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 7

[11] Young-Woon Cha, True Price, Zhen Wei, Xinran Lu, Nicholas Rewkowski, Rohan Chabra, Zihe Qin, Hyounghun Kim, Zhaoqi Su, Yebin Liu, et al. Towards fully mobile 3d face, body, and environment capture using only head-worn cameras. *IEEE transactions on visualization and computer graphics*, 24(11):2993–3004, 2018. 8

[12] Young-Woon Cha, Husam Shaik, Qian Zhang, Fan Feng, Andrei State, Adrian Ilie, and Henry Fuchs. Mobile. egocentric human body motion reconstruction using only eyeglasses-mounted cameras and a few body-worn inertial sensors. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 616–625. IEEE, 2021. 8

[13] Chen Chen, Weixing Zhu, and Tomas Norton. Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning. *Computers and Electronics in Agriculture*, 187:106255, 2021. 8

[14] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2090–2099, 2019. 7

[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*, 105:104069, 2021. 5

[16] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018. 8

[17] Mertcan Cokbas, John Bolognino, Janusz Konrad, and Prakash Ishwar. Frida: Fisheye re-identification dataset with annotations. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2022. 4

[18] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 518–533, 2018. 8

[19] Thiago L. T. da Silveira, Paulo G. L. Pinto, Jeffri Murrugarra-Llerena, and Cláudio R. Jung. 3d scene geometry estimation from 360° imagery: A survey. *ACM Comput. Surv.*, 55(4), nov 2022. 1

[20] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 7

[21] Carlos R del Blanco, Pablo Carballeira, Fernando Jaureguizar, and Narciso García. Robust people indoor localization with omnidirectional cameras using a grid of spatial-aware classifiers. *Signal Processing: Image Communication*, 93:116135, 2021. 3, 6

[22] Konstantinos K Delibasis, Spiros V Georgakopoulos, Konstantina Kottari, Vassilis P Plagianakos, and Ilias Maglogiannis. Geodesically-corrected zernike descriptors for pose recognition in omni-directional images. *Integrated Computer-Aided Engineering*, 23(2):185–199, 2016. 7

[23] Konstantinos K. Delibasis and Ilias Maglogiannis. A fall detection algorithm for indoor video sequences captured by fish-eye camera. In *2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 1–5, 2015. 8

[24] Banş Evrim Demiröz, Ismail Ari, Orhan Eroğlu, Albert Ali Salah, and Laie Akarun. Feature-based tracking on a multi-omnidirectional camera dataset. In *2012 5th International Symposium on Communications, Control and Signal Processing*, pages 1–5. IEEE, 2012. 3

[25] Julia Denecke and Christian Jauch. Verification and regularization method for 3d-human body pose estimation based on prior knowledge. *Electronic Imaging*, 33:1–8, 2021. 7

[26] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 7

[27] Zhihao Duan, Ozan Tezcan, Hayato Nakamura, Prakash Ishwar, and Janusz Konrad. Rapid: rotation-aware people detection in overhead fisheye images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 636–637, 2020. 3, 6, 8

[28] Andrea Eichenseer and André Kaup. A data set providing synthetic and real-world fisheye video sequences. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1541–1545, 2016. 4

[29] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 3, 5

[30] Dario Figueira, Matteo Taiana, Athira Nambiar, Jacinto Nascimento, and Alexandre Bernardino. The hda+ data set for research on fully automated re-identification systems. In *European Conference on Computer Vision*, pages 241–255. Springer, 2014. 3

[31] Raspberry Pi Foundation. Raspberry pi 3 model b. https://www.raspberrypi.com/products/raspberry-pi-3-model-b/, 2016. [Online; accessed 11-April-2023]. 6

[32] Daniel Fuertes, Carlos R del Blanco, Pablo Carballeira, Fernando Jaureguizar, and Narciso García. People detection with omnidirectional cameras using a spatial grid of deep learning foveatic classifiers. *Digital Signal Processing*, 126:103473, 2022. 6

[33] Nicola Garau, Niccolo Bisagno, Piotr Bródka, and Nicola Conci. Deca: Deep viewpoint-equivariant human pose estimation using capsule autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11677–11686, 2021. 7

[34] Nicola Garau, Giulia Martinelli, Piotr Bródka, Niccolò Bisagno, and Nicola Conci. Panoptop: A framework for generating viewpoint-invariant human pose estimation datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 234–242, October 2021. 5

[35] José Gaspar, Niall Winters, and José Santos-Victor. Vision-based navigation and environmental representations with an omnidirectional camera. *IEEE Transactions on robotics and automation*, 16(6):890–898, 2000. 1

[36] Spiros V Georgakopoulos, Konstantina Kottari, Kostas Delibasis, Vassilis P Plagianakos, and Ilias Maglogiannis. Pose recognition using convolutional neural networks on omni-directional images. *Neurocomputing*, 280:23–31, 2018. 7

[37] Olfa Haggui, Hamza Bayd, Baptiste Magnier, and Arezki Aberkane. Human detection in moving fisheye camera using an improved yolov3 framework. In *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2021. 6

[38] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1229–1238, 2016. 5, 7

[39] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Li Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 160–177. Springer, 2016. 7

[40] Richard I Hartley, Rajiv Gupta, and Tom Chang. Stereo from uncalibrated cameras. In *CVPR*, volume 92, pages 761–764, 1992. 7

[41] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 6

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[43] Christoph Heindl, Thomas Pönitz, Andreas Pichler, and Josef Scharinger. Large area 3d human pose detection via stereo reconstruction in panoramic cameras. *arXiv preprint arXiv:1907.00534*, 2019. 7

[44] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 7

[45] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 8

[46] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472. IEEE, 2011. 7

[47] J. Kannala and S.S. Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1335–1340, 2006. 2

[48] Sung-Tae Kim, Ming Fan, Seung-Won Jung, and Sung-Jea Ko. External vehicle positioning system using multiple fish-eye surveillance cameras for indoor parking lots. *IEEE Systems Journal*, 15(4):5107–5118, 2020. 8

[49] Konstantina N Kottari, Konstantinos K Delibasis, and Ilias G Maglogiannis. Real-time fall detection using uncalibrated fisheye cameras. *IEEE Transactions on Cognitive and Developmental Systems*, 12(3):588–600, 2019. 8

[50] Varun Ravi Kumar, Ciarán Eising, Christian Witt, and Senthil Yogamani. Surround-view fisheye camera perception for automated driving: Overview, survey & challenges.

IEEE Transactions on Intelligent Transportation Systems, pages 1–22, 2023. 1

[51] Varun Ravi Kumar, Marvin Klingner, Senthil Yogamani, Stefan Milz, Tim Fingscheidt, and Patrick Mader. Syndistnet: Self-supervised monocular fisheye camera distance estimation synergized with semantic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 61–71, January 2021. 1

[52] Olivier Laurendin, Sébastien Ambellouis, Anthony Fleury, Ankur Mahtani, Sanaa Chafik, and Clément Strauss. Hazardous events detection in automatic train doors vicinity using deep neural networks. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7. IEEE, 2021. 1, 4, 8

[53] Junnan Li, Jianquan Liu, Wong Yongkang, Shoji Nishimura, and Mohan Kankanhalli. Weakly-supervised multi-person action recognition in 360° videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 508–516, 2020. 4, 8

[54] Jia Li, Kaiwen Yu, Yifan Zhao, Yu Zhang, and Long Xu. Cross-reference stitching quality assessment for 360° omnidirectional images. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2360–2368, New York, NY, USA, 2019. Association for Computing Machinery. 1

[55] Mingyang Li, Yi Zhou, Ming Meng, Yuehua Wang, and Zhong Zhou. 3d room reconstruction from a single fisheye image. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. 1, 8

[56] Shengye Li, M. Ozan Tezcan, Prakash Ishwar, and Janusz Konrad. Supervised people counting using an overhead fisheye camera. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2019. 1, 3, 5, 6

[57] Daniele Liciotti, Marina Paolanti, Emanuele Frontoni, Adriano Mancini, and Primo Zingaretti. *Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration*, pages 1–11. Springer International Publishing, Cham, 2017. 5

[58] Daniele Liciotti, Marina Paolanti, Emanuele Frontoni, and Primo Zingaretti. People detection and tracking from an rgb-d camera in top-view configuration: review of challenges and applications. In *New Trends in Image Analysis and Processing–ICIAP 2017: ICIAP International Workshops, WBICV, SSPandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, September 11-15, 2017, Revised Selected Papers 19*, pages 207–218. Springer, 2017. 1

[59] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 8

[60] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[61] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. 5, 7

[62] Zhangchi Lu., Mertcan Cokbas., Prakash Ishwar., and Janusz Konrad. Estimating distances between people using a single overhead fisheye camera with application to social-distancing oversight. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 528–535. INSTICC, SciTePress, 2023. 4

[63] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 8

[64] Quan Nguyen Minh, Bang Le Van, Can Nguyen, Anh Le, and Viet Dung Nguyen. Arpd: Anchor-free rotation-aware people detection using topview fisheye camera. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2021. 1, 6

[65] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015. 6

[66] S.K. Nayar. Catadioptric omnidirectional camera. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 482–488, 1997. 1

[67] Thanh Binh Nguyen, Van Tuan Nguyen, Sun-Tae Chung, and Seongwon Cho. Real-time human detection under omni-directional camera based on cnn with unified detection and agmm for visual surveillance. *Journal of Korea Multimedia Society*, 19(8):1345–1360, 2016. 1, 5, 6

[68] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1273–1283, 2019. 8

[69] Viet Dung Nguyen, Phuc Ngoc Pham, Xuan Bach Nguyen, Thi Men Tran, and Minh Quan Nguyen. Incorporation of panoramic view in fall detection using omnidirectional camera. In *The International Conference on Intelligent Systems & Networks*, pages 313–318. Springer, 2021. 8

[70] Kazuhiro Otsuka, Shoko Araki, Dan Mikami, Kentaro Ishizuka, Masakiyo Fujimoto, and Junji Yamato. Realtime meeting analysis and 3d meeting viewer based on omnidirectional multimodal sensors. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 219–220, 2009. 1

[71] G Dias Pais, Tiago J Dias, Jacinto C Nascimento, and Pedro Miraldo. Omnidrl: Robust pedestrian detection using

deep reinforcement learning on omnidirectional cameras. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4782–4789. IEEE, 2019. 6, 7

[72] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2272–2280, January 2021. 1

[73] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 5

[74] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 5

[75] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 5

[76] Naveed ur Rehman. Hemispherical photographs: A review of acquisition methods and applications in the context of urban energy and environment assessments. *ASME Open Journal of Engineering*, 1:010801, 2022. 2

[77] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6, 7

[78] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 7

[79] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2

[80] Tobias Scheck, Ana Grassi, and Gangolf Hirtz. Unsupervised domain adaptation from synthetic to real images for anchorless object detection. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,*, pages 319–327. INSTICC, SciTePress, 2021. 7

[81] T. Scheck, R. Seidel, and G. Hirtz. Learning from theodore: A synthetic omnidirectional top-view indoor dataset for deep transfer learning. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 932–941, Los Alamitos, CA, USA, mar 2020. IEEE Computer Society. 4, 5, 7

[82] Miriam Schönbein and Andreas Geiger. Omnidirectional 3d reconstruction in augmented manhattan worlds. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 716–723, 2014. 1

[83] Roman Seidel, André Apitzsch, and Gangolf Hirtz. Improved person detection on omnidirectional images with non-maxima supression. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 474–481. INSTICC, SciTePress, 2019. 5, 6

[84] Roman Seidel, André Apitzsch, Jingrui Yu, Julian Seuffert, Norbert Nestler, Danny Heinz, Anne Goy, and Gangolf Hirtz. Auxilia: Nutzerzentriertes assistenz-und sicherheitssystem zur unterstützung von menschen mit demenz auf basis intelligenter verhaltensanalyse. In *Innteract Conference*, volume 2018, page 48, 2018. 1, 7

[85] Ahmed Rida Sekkat, Yohan Dupuis, Varun Ravi Kumar, Hazem Rashed, Senthil Yogamani, Pascal Vasseur, and Paul Honeine. Synwoodscape: Synthetic surround-view fisheye camera dataset for autonomous driving. *IEEE Robotics and Automation Letters*, 7(3):8502–8509, 2022. 1

[86] Julian Bruno Seuffert, Ana Cecilia Perez Grassi, Tobias Scheck, and Gangolf Hirtz. A Study on the Influence of Omnidirectional Distortion on CNN-based Stereo Vision. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2021, Volume 5: VISAPP*, pages 809–816, Online Conference, 2 2021. SciTePress. 5

[87] Roman Siedel, Tobias Scheck, Ana C Perez Grassi, Julian B Seuffert, André Apitzsch, Jingrui Yu, Norbert Nestler, Danny Heinz, Lars Lehmann, Anne Goy, et al. Contactless interactive fall detection and sleep quality estimation for supporting elderly with incipient dementia. *Current Directions in Biomedical Engineering*, 6(3):388–391, 2020. 8

[88] Karen Stephen, Jianquan Liu, and Vivek Barsopia. A hybrid two-stream approach for multi-person action recognition in top-view 360° videos. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3418–3422. IEEE, 2021. 8

[89] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems*, 2021. 8

[90] Masato Tamura, Shota Horiguchi, and Tomokazu Murakami. Omnidirectional pedestrian detection by rotation invariant training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1989–1998, 2019. 3, 6

[91] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 503–512, 2022. 3, 6

[92] Denis Tome, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando De la Torre. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 7

[93] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an

hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7728–7738, 2019. 7

[94] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International journal of computer vision*, 101(1):184–204, 2013. 3

[95] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 7

[96] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt. Estimating egocentric 3d human pose in the wild with external weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13157–13166, June 2022. 8

[97] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11500–11509, 2021. 7

[98] Tsaipei Wang, Yun-Yi Hsieh, Fong-Wen Wong, and Yi-Fu Chen. Mask-rcnn based people detection using a top-view fisheye camera. In *2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 1–4. IEEE, 2019. 6

[99] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022. 8

[100] Thaddäus Wiedemer, Stefan Wolf, Arne Schumann, Kaisheng Ma, and Jürgen Beyerer. Few-shot supervised prototype alignment for pedestrian detection on fisheye images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4142–4153, June 2022. 6

[101] N. Winters, J. Gaspar, G. Lacey, and J. Santos-Victor. Omni-directional vision for robot navigation. In *Proceedings IEEE Workshop on Omnidirectional Vision (Cat. No.PR00704)*, pages 21–28, 2000. 1

[102] Changhee Won, Hochang Seok, Zhaopeng Cui, Marc Pollefeys, and Jongwoo Lim. Omnislam: Omnidirectional localization and dense mapping for wide-baseline multi-camera systems. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 559–566. IEEE, 2020. 1

[103] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5):2093–2101, 2019. 7

[104] Marie Yahiaoui, Hazem Rashed, Letizia Mariotti, Ganesh Sistu, Ian Clancy, Lucie Yahiaoui, Varun Ravi Kumar, and Senthil Yogamani. Fisheyemodnet: Moving object detection on surround-view cameras for autonomous driving. *arXiv preprint arXiv:1908.11789*, 2019. 1

[105] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Conference on Fairness, Accountability, and Transparency*, 2020. 2

[106] Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sumanth Chennupati, Sanjaya Nayak, Saquib Mansoor, Xavier Perrotton, and Patrick Perez. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[107] Jingrui Yu, Roman Seidel, and Gangolf Hirtz. Omnipd: One-step person detection in top-view omnidirectional indoor scenes. *Current Directions in Biomedical Engineering*, 5(1):239–244, 2019. 3, 5, 6

[108] Dongxu Zhao, Zhen Wei, Jisan Mahmud, and Jan-Michael Frahm. Egoglass: Egocentric-view human pose estimation from an eyeglass frame. In *2021 International Conference on 3D Vision (3DV)*, pages 32–41. IEEE, 2021. 8