IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 April 2024, accepted 27 April 2024, date of publication 1 May 2024, date of current version 16 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3395786

APPLIED RESEARCH

Improving Passenger Detection With Overhead Fisheye Imaging

DIMITRIS TSIKTSIRIS[1,2], ANTONIOS LALAS[1], MINAS DASYGENIS[2], AND KONSTANTINOS VOTIS[1]

[1]Information Technologies Institute, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece
[2]Department of Electrical and Computer Engineering, University of Western Macedonia, 50100 Kozani, Greece

Corresponding author: Dimitris Tsiktsiris (tsiktsiris@iti.gr)

This work was supported by European Union's Horizon Europe Research and Innovation Program "Advancing Sustainable User-Centric Mobility with Automated Vehicles (ULTIMO)" under Grant Agreement 101077587.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the CERTH Ethical Committee.

ABSTRACT Detecting passengers within overhead, fisheye images presents a unique set of challenges. Traditional approaches rely on radially-aligned bounding boxes based on the assumption that people are consistently oriented along the image radius. This assumption simplifies the detection process but introduces limitations in terms of flexibility and detection accuracy. Additionally, these methods often require extensive pre and post-processing, significantly increasing the computational complexity. We propose an innovative, end-to-end, rotation-aware detection framework specifically designed for the accurate detection of passengers using angle-oriented bounding boxes. This study investigates a fully convolutional neural network (CNN) that performs direct orientation regression of each bounding box, enhanced by a scale and angle loss function that effectively accounts for the periodicity of angles, ensuring accurate and robust bounding box orientation predictions. Moreover, we present a new dataset tailored to in-cabin passenger detection and counting. Our experimental results show an improvement of 5.3% in average precision, compared to state-of-the-art methods in overhead people detection. Finally, we demonstrate results from real vehicle experiments in Copenhagen and Geneva, highlighting the importance of this work for public transport operators.

INDEX TERMS Artificial intelligence, edge computing, autonomous vehicles, computer vision, in-cabin monitoring, passenger counting.

I. INTRODUCTION

Automated passenger counting (APC) is a crucial component for the successful integration of autonomous vehicles (AVs) in public transportation. APC systems enable public transport operators (PTOs) to accurately monitor the number of passengers on board in real time, which is instrumental in optimizing resource allocation, such as the allocation of vehicles to routes with higher passenger demand, and helping to ensure that AVs are efficiently utilized, reducing unnecessary energy consumption and operational costs. Moreover, knowing the exact number of passengers onboard is crucial for maintaining safety standards, as overcrowding can lead to safety hazards and discomfort for passengers. By employing APC, AVs can prevent overloading and ensure that passengers travel comfortably and securely. Additionally, the data collected by APC systems provide valuable insights into passenger behavior and preferences, offering PTOs the ability to make decisions about route planning, scheduling, and service adjustments, improving the overall quality of public transportation services. Finally, AVs equipped with APC systems can provide real-time information about passenger occupancy through onboard displays or mobile apps, also allowing passengers to make informed decisions

The associate editor coordinating the review of this manuscript and approving it for publication was Jie Gao.

FIGURE 1. YOLO-V5 inference on overhead fisheye camera image.

about their travel plans, reducing wait times and enhancing their overall experience.

A typical APC system consists of hardware sensors, typically video cameras, with computer vision algorithms [1], [2], [3] and a setup that usually involves the lateral positioning of a wide field of view camera above the area of interest, with the use of multiple such cameras to cover larger areas. An intriguing alternative to this conventional setup would be the utilization of a singular fisheye lens boasting a 360-degree field of view (FoV). Although, the current problem with such setups lies in the already existing detection algorithms that were originally designed for side perspective, standard-lens images, thereby often struggling when applied to overhead 360-degree images, primarily due to the distinctive circular geometry and barrel-shaped distortions characteristic of the latter [4], [5].

In conventional images, where standing individuals are typically depicted in a vertical orientation, detection algorithms that recognize bounding boxes aligned with the image axes, such as You Only Look Once (YOLO) [6], Single-Shot Detector (SSD) [7], and Region-based Convolutional Neural Network (R-CNN) [8], perform admirably in such scenarios. However, these same algorithms encounter significant challenges when applied to overhead fisheye images [9], frequently failing to detect individuals not in upright postures, as illustrated in Fig. 1. In these overhead images, standing people are often situated along the image radius due to the camera's overhead placement, necessitating the use of rotated bounding boxes.

To address this need for rotation-aware detection, various YOLO-based people detection algorithms have been recently proposed [9], [10], [11], [12], [13], [14], each tackling the radial geometry differently. For instance, a notable approach involves rotating the image in small 15-degree increments, followed by applying YOLO to the central upper part of the image, where people are generally upright, and then performing post-processing to eliminate redundancies [9]. However, this approach demands multiple YOLO applications. On that note, another recent approach [13] employs rotated bounding boxes to align people with the radial axis of the image, which can typically be less effective in identifying individuals in non-upright positions, as depicted in Fig. 1.

In this paper, a novel end-to-end detection algorithm for passengers inside autonomous vehicles is designed specifically for overhead fisheye images based on a single-stage CNN, extending the architecture originally introduced in YOLO [6], [15], [16]. Apart from predicting the centroid and size of the bounding boxes, we also incorporate the use of an angle-aware loss function extending scale-invariant regression loss for angle prediction, thus enabling the precise determination of the orientation of each bounding box without adding computational complexity. Being an end-to-end solution, the proposed method allows for training or fine-tuning using annotated fisheye images, with fine-tuning from models trained on standard images showing significant performance improvements. In addition, another noteworthy advantage of this proposed work, driven by its focus on passenger detection, is the adoption of single-class object detection, replacing the common regression-based loss function found in multi-class object detection algorithms [6], [7], [17], [18]. Remarkably, the inference speed closely matches that of YOLO, as it processes each image only once without requiring pre- or post-processing steps.

The remainder of this work is structured as follows. Section II offers a survey of the most recent bibliography on already existing overhead fisheye imaging solutions from the state-of-the-art. In Section III, the overall architecture of our proposed overhead detection framework is presented, while in Section IV, a comprehensive analysis of the oriented bounding box regression method is offered. Finally, Sections V and VII offer the experimental results of this work, derived after the implementation of the proposed detection framework, and summarize our conclusions, respectively.

II. RELATED WORK

The exploration of both people and object detection techniques across various imaging modalities has led to significant advancements in three distinct areas by employing side-view, standard-lens cameras, where traditional and deep learning methods excel in people detection through feature analysis and direct bounding box regression; utilizing rotated bounding boxes by adapting object detection frameworks to accommodate orientation gradations; and, lately, by employing overhead, fisheye imaging, thus prompting the development of specialized algorithms. This section presents existing people and object detection approaches from the state-of-the-art that fit within these three categories, highlighting the evolution of detection methodologies tailored to the specific characteristics of each imaging approach.

A. SIDE-VIEW, STANDARD-LENS DETECTION

Detecting individuals with conventional side-view cameras leveraging standard lenses involves well-known methods including histograms of oriented gradients (HOG) [19] and aggregate channel features (ACF) [20]. The advent of deep learning has considerably enhanced the efficiency of identifying objects and individuals [6], [7], [8], [17], [18], [21], leading to a classification of two main types, namely

two-stage and one-stage methods. More specifically, the two-stage approach, including R-CNN and its variants [8], [17], [18], employs a Region Proposal Network (RPN) to predict a region of interest (ROI), followed by precise bounding box adjustments. On the other hand, one-stage methods, such as SSD [7], [21] and YOLO [6], [15], [16] variations, operate as standalone RPNs that directly deduce bounding boxes from the input image via CNNs. Recently, the focus has shifted towards both fast one-stage detectors [22], [23] and anchor-free detectors [24], [25].

B. ROTATED BOUNDING BOXES DETECTION

Another research field that gained interest is the study of detecting objects with rotated bounding boxes, especially for text recognition and aerial imagery analysis [26], [27], [28], [29]. In this context, the rotated RPN (RRPN) algorithm [26] introduces a dual-stage detection process incorporating rotated anchors and a unique rotated ROI layer, while the RoI-Transformer [27] goes one step further by initially setting a horizontal ROI and then adapting it into a rotated form. R3Det [28] is another innovative algorithm with a single-stage detector that includes a refinement layer to address alignment discrepancies, a challenge common in single-stage frameworks. Alternately, Nosaka et al. [30] employ orientation-sensitive convolutional layers [14] for orientation adjustment, as well as a smooth L1 loss for angle correction. All of these aforementioned methodologies typically represent bounding boxes with a five-element vector, accounting for symmetry in their orientation to refine loss calculation and addressing the limitations of traditional regression losses when estimating closely aligned predictions with the actual position and orientation. RSDet [29] addresses this by introducing a modulated rotation loss.

C. OVERHEAD, FISHEYE DETECTION

Detecting individuals in overhead fisheye imagery is a new yet developing field with limited existing research. Traditional detection methods, including HOG and local binary patterns (LBP), have been adapted for the unique distortions of fisheye lenses with slight adjustments to account for fisheye geometry [10], [11], [12], [31]. More specifically, new techniques involve rotating fisheye imagery in increments to capture and analyze features from specific image sections using classifiers, such as support vector machines (SVM) [10]. Additionally, adaptations have been made to adjust feature extraction methods to the fisheye perspective for accurate person identification [12].

Additional approaches also include deep learning techniques that propose CNN modifications to accommodate the fisheye distortion by adjusting CNNs into rotation-invariant approaches and applying them to specifically processed or transformed versions of the image data to align with typical image orientations [6], [9], [32]. Such approaches involve preprocessing by dewarping or rotating the images, followed by sophisticated post-processing to refine detection accuracy and reduce redundancy in detection instances. The work by Duan et al. [33] features a similar architecture, containing a backbone, an FPN and a detection head for angle prediction.

D. PURPOSE AND SCOPE

This study introduces a novel approach by integrating an angle loss function for precise bounding box orientation prediction and revising the rotated bounding box representation to address inherent symmetry issues. This enhances the accuracy and efficiency of passenger detection in fisheye images from on-board footage. More specifically, we introduce a new loss function, taking into account the scale and the angles of the bounding boxes and focusing on the conditions inside the vehicle's cabin. Moreover, we present a dataset for passenger detection, specifically tailored for automated passenger counting in autonomous vehicles. Finally, we evaluate our proposed passenger counting method through qualitative and quantitative analysis of a real-world installation. The results indicate that our method outperforms traditional regression loss methods without introducing additional computational complexity.

III. METHODOLOGY

Similar to one-stage detectors, the proposed overhead imaging model contains a Backbone network, a Feature Pyramid Network (FPN) and a bounding-box regression network as a detection head, with each component specifically designed to process the input image and extract features at multiple scales, resulting in the prediction of bounding boxes and class categories for objects within the image, as depicted in Fig. 2. The choice for a single-stage approach was made to satisfy the need for real-time processing in our solution, a resource-constrained embedded device with low power requirements. Single-stage CNNs offer significantly faster inference speeds due to their simple design, compared to two-stage detectors [34].

A. BACKBONE NETWORK

The backbone of the network is responsible for the initial feature extraction and is typically a pre-trained CNN such as ResNet, VGG, or a similar architecture. More specifically, given an input image I, the network produces a set of multi-dimensional feature maps $\{P_k\}_{k=1}^{3}$ at different scales, denoted as P_1, P_2, and P_3 for high, medium, and low resolutions, respectively, as follows:

$$P_k = \text{Backbone}(I), \quad k \in \{1, 2, 3\} \tag{1}$$

B. FEATURE PYRAMID NETWORK

The FPN enhances the backbone's feature maps by integrating high-level semantic information from deep layers with spatial information from earlier layers based on eq. (2). This choice was made due to the significant variations in object sizes due to the distortion of the wide-angle fisheye lenses. FPN's multi-scale feature representations allow for effectively detecting both passengers in the center, who

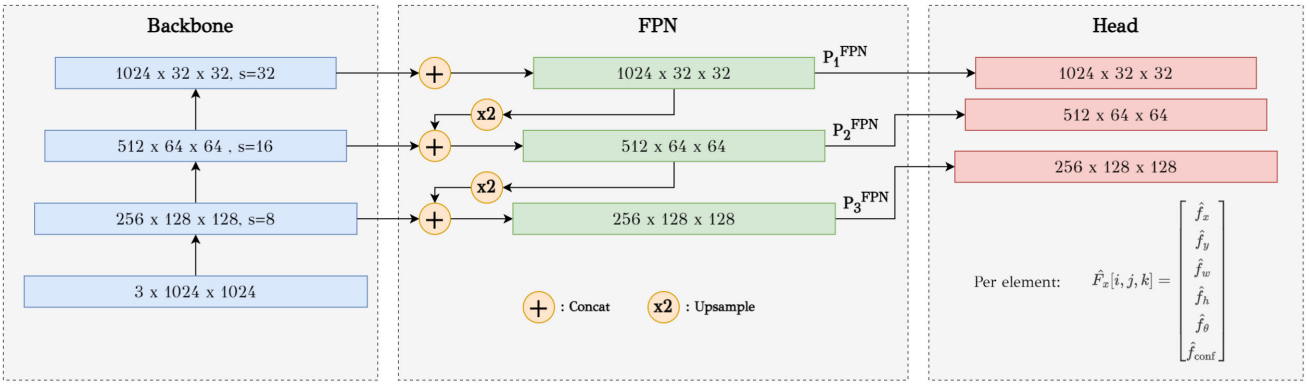


FIGURE 2. Model architecture. Each arrow represents multiple convolutional layers, and the colored rectangles represent multi-dimensional matrices, i.e., feature maps, whose dimensions correspond to an input image of size $h \times w = 1,024 \times 1,024$.

appear larger, and those who appear much smaller due to the peripheral distortion. This is achieved by upsampling spatially coarser, but semantically stronger, feature maps from higher pyramid levels and merging them with lower-level maps through element-wise addition.

$$\{P_k^{\text{fpn}}\} = \text{FPN}(\{P_k\}), \quad k \in \{1, 2, 3\} \tag{2}$$

C. DETECTION HEAD

Post feature enhancement by the FPN, the detection head predicts the transformed bounding boxes and class scores by applying a separate CNN to each FPN feature vector to generate a transformed bounding box predictions $\tilde{T}_k$ and an objectness score, which are relatively parameterized to anchor pre-defined boxes at different scales and aspect ratios, as follows:

$$\tilde{t}_{x,k} = s_k \cdot (i + \text{Sigmoid}(f_{x,k})) \tag{3}$$
$$\tilde{t}_{y,k} = s_k \cdot (j + \text{Sigmoid}(f_{y,k})) \tag{4}$$
$$\tilde{t}_{w,k} = w_k^{\text{anchor}} \cdot \exp(f_{w,k}) \tag{5}$$
$$\tilde{t}_{h,k} = h_k^{\text{anchor}} \cdot \exp(f_{h,k}) \tag{6}$$
$$\tilde{o}_k = \text{Sigmoid}(f_{o,k}), \tag{7}$$

where i and j represent the center of the anchor box, s_k is the stride of the feature map at scale k, and $f_{x,k}, f_{y,k}, f_{w,k}, f_{h,k}, f_{o,k}$ are the outputs of the CNN applied to the feature vector from the FPN. The variables w_k^{anchor} and h_k^{anchor} represent the width and height of the k-th anchor box, and Sigmoid is the logistic function applied to constrain the outputs to a range between 0 and 1. The objectness score $\tilde{o}_k$ denotes the probability that an object is present within the predicted bounding box. The training loss function of the network combines the regression loss for the bounding box coordinates and a classification loss for the objectness score.

IV. ROTATION-AWARE BOUNDING BOX REGRESSION

The rotation-aware bounding box regression is a critical component of the proposed method, enabling the detection of people in overhead fisheye images with a high degree of accuracy in both position and orientation by incorporating a novel angle prediction mechanism that accounts for the unique properties of angles as cyclic quantities. It involves the prediction of a bounding box's orientation, along with its center and size, which are normalized relative to the feature map dimensions as follows:

$$\mathcal{L}_{reg} = \sum_i^N \left(\lambda_{coord} \text{Scale}_{L1}(b_i, \hat{b}_i) + \lambda_{angle} \text{Periodic}_{L1}(\theta_i, \hat{\theta}_i) \right) \tag{8}$$

Here, $\mathcal{L}_{reg}$ represents the regression loss, b_i the predicted bounding box, $\hat{b}_i$ the ground truth box, θ_i the predicted orientation, and $\hat{\theta}_i$ the ground truth orientation. The terms Scale_{L1} and Periodic_{L1} are the scale L1 loss and periodic L1 loss, respectively, with λ_{coord} and λ_{angle} as the balancing weights. This allows the network to effectively learn the orientation of objects, taking into consideration the periodic nature of the angle, thus providing a more robust and accurate object detection in fisheye images.

A. SCALE-INVARIANT LOSS FUNCTION

To address scale variance in object detection, particularly for fisheye images, we propose the integration of a scale-invariant term in the loss function that aims to stabilize the learning across different object sizes, a common challenge in fisheye image datasets due to perspective distortion. The scale-invariant loss term L_{scale} could be formulated as follows:

$$L_{scale} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{w_i h_i}} L_{obj}(o_i, \hat{o}_i), \tag{9}$$

where N is the number of objects, w_i and h_i are the width and height of the bounding box of the i-th object, L_{obj} is the objectness loss for the i-th object, o_i is the ground truth objectness score, and $\hat{o}_i$ is the predicted objectness score.

B. ANGLE LOSS FUNCTION

The angle loss is crucial for the model's ability to learn the orientation of objects, a fundamental aspect when dealing with fisheye images. It ensures accurate angle predictions for bounding boxes by incorporating binary cross-entropy (*BCE*) for foreground-background classification with a specialized term for angle regression as follows:

$$\begin{aligned}\mathcal{L}_{angle} = &\sum \text{BCE}(\sigma(t_{cls}), y_{cls}) \\ &+ \sum \text{BCE}(\sigma(t_{obj}), y_{obj}) \\ &+ \sum_{pos} \lambda_\theta \mathcal{L}_{periodic}(\theta_p, \theta_g)\end{aligned} \tag{10}$$

where $\mathcal{L}_{angle}$ is the composite loss for classification and angle prediction, σ the sigmoid function, t_{cls} the class logits, t_{obj} the objectness logits, y_{cls} and y_{obj} the class and objectness labels, λ_θ the weight for angle loss, $\mathcal{L}_{periodic}$ the periodic angle loss, θ_p the predicted angle, and θ_g the ground truth angle. This harmonized loss function facilitates the network to learn not only the presence of an object but also its precise rotational alignment.

C. PERIODIC ANGLE PREDICTION LOSS

The periodic angle prediction loss mitigates angle discontinuity by employing a periodic loss. The network learns to effectively predict angles in a rotation-invariant manner, which is critical for maintaining consistency when angles form a full circle.

$$\mathcal{L}_{periodic}(\theta_p, \theta_g) = \min_{k \in \mathbb{Z}} f(\theta_p - \theta_g + 2\pi k) \tag{11}$$

Here, θ_p is the predicted angle, θ_g is the ground truth angle, and f is a distance metric, such as the $L2$ norm. The loss function ensures smooth transitions across the angle boundary, effectively treating angles 2π radians apart as equivalent. The minimization over k accounts for the multiple equivalent representations of the same angle due to periodicity, thereby encouraging the network to learn angle predictions that are robust against rotational variances.

V. EXPERIMENTAL RESULTS

In this section, the outcomes from our proposed research are presented, in order to highlight key observations, trends, as well as insights gained throughout the experimentation process.

A. DATASET

While numerous datasets exist for detecting people from overhead fisheye images, they either lack annotations with rotated bounding boxes [35] or have limitations in terms of the number of frames and individuals [9]. HABBOF and CEPDOF datasets [33] are fully-featured datasets, suitable for our experiments, containing samples from an office environment. To adapt better to in-cabin conditions in the autonomous vehicle, we collected an additional dataset entitled "Autonomous Vehicles Overhead Fisheye (AVOF)". For the evaluation of our proposed method and its comparison with previous state-of-the-art techniques, the HABBOF, CEPDOF and AVOF datasets were utilized, with Table 1 containing various statistics for each one. As seen, the AVOF dataset contains a significantly larger number of frames and human objects, featuring challenging scenarios in the vehicle's cabin, such as crowded conditions, occlusion scenarios, light variations and low-light conditions, as depicted in Fig. 3, which are absent in the other datasets. In addition, both AVOF and CEPDOF are spatio-temporally annotated, ensuring that bounding boxes of the same individual carry consistent IDs across consecutive frames, rendering them suitable for vision tasks involving overhead fisheye perspective, such as passenger tracking and re-identification. Informed consent was obtained for every human subject. To increase dataset diversity and model robustness, a data augmentation pipeline performs geometric transformations like rotation (up to 15 degrees with a probability of 0.5) and horizontal flipping with a probability of 0.5. Additionally, brightness and contrast are randomly adjusted within a specified range, with a probability ratio of 0.7 and image normalization is applied to standardize the input data.

TABLE 1. Statistics of our new AVOF dataset in comparison with existing overhead fisheye image datasets. The dataset contain challenging scenarios in the vehicle's cabin, such as crowded conditions, occlusion scenarios, light variations and low-light conditions.

Dataset	Resolution	Segs	# avg/max	Frames	FPS
HABBOF	2048	4	3.5/5	5.837	30
CEPDOF	1080-2048	8	6.8/13	25.504	1-10
AVOF	2048	32	3.8/9	14.400	15
Scenarios	**AVOF**				
Stationary	2048	7	6.0	3.150	15
Moving	2048	11	6.0	4.950	15
Crowded	2048	6	10.8	2.700	15
Edge Cases	2048	4	5.5	1.800	15
Night (IR)	2048	5	6.8	2.250	15

B. PERFORMANCE METRICS

Following the MS COCO challenge [36], average precision (AP) was commonly utilized as one of the evaluation criteria, especially the area under the Precision-Recall curve. However, due to the inherent ambiguity in ground truth annotations, we focus on AP at IoU = 0.5 (AP_{50}), as even with a perfect algorithm the IoU might be relatively low. The reason behind this is the multiple bounding boxes that can exist at various angles for the same individual and a single choice of human annotator as the ground truth. Apart from AP, F-measure was also incorporated at a constant confidence threshold ($\hat{b}_{conf} = 0.3$) as another performance indicator, which, for a specific $\hat{b}_{conf}$ value, corresponds to a particular point on the Precision-Recall curve. The choice

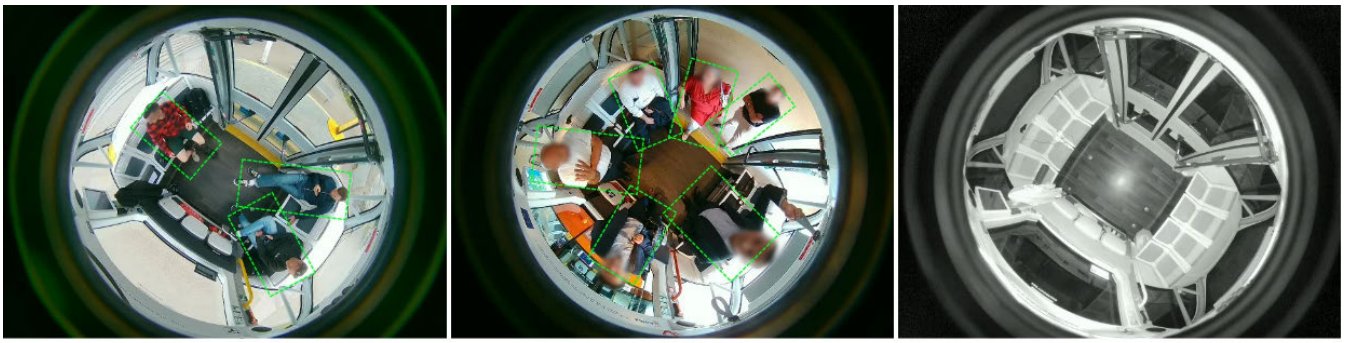


FIGURE 3. AVOF samples for three different scenarios: (a) an average number of passengers onboard; (b) a crowded scenario; and (c) with no passengers inside the cabin.

TABLE 2. Performance comparison of various methods on the HABBOF, CEPDOF and AVOF datasets on RTX 4090. Numbers in parentheses indicate the input resolution (multiplied by a power of two).

		HABBOF				CEPDOF				AVOF			
Method	**FPS**	**AP_{50}**	**P**	**R**	**F**	**AP_{50}**	**P**	**R**	**F**	**AP_{50}**	**P**	**R**	**F**
(608) [13]	51.2	87.3	0.970	0.827	0.892	61.0	0.969	0.526	0.634	62.5	0.945	0.810	0.873
AA (1,024) [9]	3.2	87.7	0.922	0.867	0.892	73.9	0.896	0.638	0.683	75.0	0.950	0.820	0.881
AB (1,024) [9]	1.7	93.7	0.881	0.935	0.907	76.9	0.884	0.694	0.743	77.5	0.955	0.825	0.886
RAPiD (608) [33]	**52.5**	97.3	0.984	0.935	0.958	82.4	0.970	0.827	0.892	85.0	0.968	0.850	0.906
RAPiD (1,024) [33]	27.7	**98.1**	0.975	0.963	0.969	85.8	0.970	0.827	0.892	87.0	0.972	0.855	0.911
Proposed (1,024)	29.1	97.9	0.960	0.931	0.951	**86.1**	0.978	0.965	0.971	**92.3**	0.960	0.940	0.949

of AP_{50} was made to favor detections that are acceptable in a practical APC application, without demanding perfect alignment. F-measure at 0.3 provides an acceptable trade-off between precision and recall, avoiding false negatives and minimizing false positives. Combined, these metrics offer a robust evaluation of the model's ability to accurately detect passengers under the specific challenges presented by the overhead fisheye perspective.

C. QUANTITATIVE RESULTS

We initiate the training process on MS COCO 2017 training images [36] for 120,000 iterations, followed by fine-tuning the network on single or multiple datasets from Table 1 for 10,000 iterations, with each iteration comprised of 112 images. During training on COCO images, the network weights are updated using Stochastic Gradient Descent (SGD) with a step size of 0.0005, a momentum of 0.9 and a 0.0003 weight decay. The learning rate is adjusted via a decay mechanism, reduced by a factor of 10 after every 30,000 iterations without improvement in validation loss for optimal convergence.

For the datasets listed in Table 1, the standard SGD was utilized with a step size of 0.0001, while rotation, flipping, shifting, resizing, and color augmentation techniques were also applied during both training stages. All results presented here are based on a single run of training and inference. The training was conducted on a system with an Intel Core i9-9900K CPU @ 3.60GHz, 64 GB of system RAM and a single NVIDIA RTX 4090 GPU with 24GB of VRAM.

Table 2 provides a comparative analysis of our method with other competing algorithms. To evaluate AA and AB algorithms from Li et al. [9], we utilize the authors' publicly-available implementation. Furthermore, given the absence of a predefined train-test split in these three datasets, a cross-validation of our method was conducted, highlighting the use of two datasets for training and the remaining one for testing, repeated so each dataset is included once as the test set.

For instance, our method is trained on HABBOF and AVOF and tested on CEPDOF, and vice versa for other transformations. As neither approach from Li et al. [9] nor Tamura et al. [13] is designed to be trained on rotated bounding boxes, for the purposes of this work, they are both trained solely on the COCO dataset, as described in their respective papers. Moreover, Tamura et al. employed a top-view standard-lens image dataset called DPI-T [37] for training, in addition to the COCO dataset; however, this dataset is currently inaccessible and thus cannot be used in this study.

Table 2 provides a detailed performance comparison of various methods evaluated on three different datasets: HABBOF, CEPDOF, and AVOF. The evaluation is conducted on an NVIDIA RTX 4090 graphics card, which is a high-end hardware platform for deep learning and computer vision tasks. The performance metrics include the Average Precision

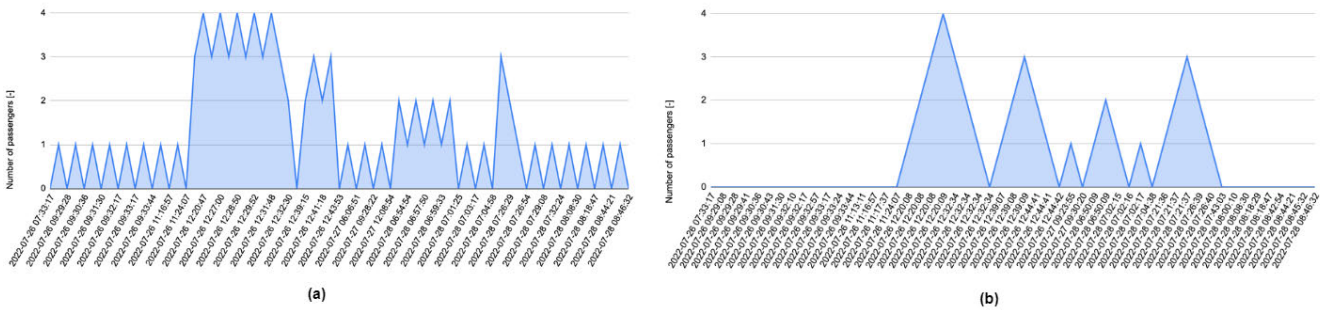
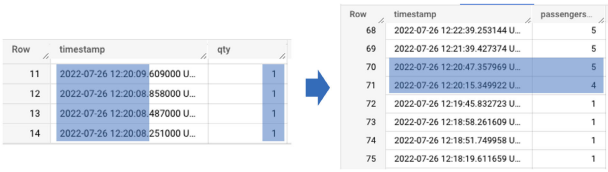


FIGURE 4. Passenger count over time received from: (a) the automated data stream; and (b) the driver's (manual counting) data stream.



Row	timestamp	qty
11	2022-07-26 12:20:09.609000 U...	1
12	2022-07-26 12:20:08.858000 U...	1
13	2022-07-26 12:20:08.487000 U...	1
14	2022-07-26 12:20:08.251000 U...	1

Row	timestamp	passengers...
68	2022-07-26 12:22:39.253144 U...	5
69	2022-07-26 12:21:39.427374 U...	5
70	2022-07-26 12:20:47.357969 U...	5
71	2022-07-26 12:20:15.349922 U...	4
72	2022-07-26 12:19:45.832723 U...	1
73	2022-07-26 12:18:58.261609 U...	1
74	2022-07-26 12:18:51.749958 U...	1
75	2022-07-26 12:18:19.611659 U...	1

FIGURE 5. Data from BigQuery: Four passengers are getting on the shuttle. In the right-side the manual count from the operator's app is depicted with data points received when state changes (button is pushed in the shuttle). In the left side the data points are received continuously from our automated method. The count includes the on-board operator.

at an IoU threshold of 50% (AP_{50}), Precision (P), Recall (R), and F1-Score (F), alongside the frame rate measured in frames per second (FPS), which indicates the inference speed of each method. These metrics collectively offer insights into the accuracy, efficiency, and speed of the evaluated methods under different resolution settings, denoted in parentheses next to each method's name, indicating the input resolution scaled by a power of two. A confidence threshold of 0.3 is used for all methods to calculate Precision, Recall, and F-measure, with test results demonstrating that our method achieves the best performance in CEPDOF and AVOF and the fastest execution speed at a resolution of 1024 × 1024 among all tested methods. The RAPiD (608) method achieves the highest FPS of 52.5 at a lower resolution of 608, making it an attractive option for real-time applications. On the other hand, the RAPiD (1,024) method showcases the best AP_{50} performance on the HABBOF dataset with a score of 98.1% at a lower frame rate of 27.7 FPS. This trade-off between accuracy and speed is a common challenge in the design and implementation of object detection systems. The proposed method, exhibits a balanced performance across all datasets, achieving a nearly top AP_{50} on HABBOF (97.9%) and the highest scores on CEPDOF (86.1% AP_{50}) and AVOF (92.3% AP_{50}). Notably, our method's performance at HABBOF is slightly lower than RAPiD (1,024), where human objects appear in an upright pose (movement), a significant observation since people walking or standing typically exhibit radial orientation in overhead fisheye images, Our method also demonstrates a high frame rate of 29.1 FPS, indicating its efficiency and suitability for applications requiring both high accuracy and real-time processing.

The system was evaluated on automated minibuses in Copenhagen and Geneva. The solution was installed on NAVYA autonomous vehicles, featuring a NVIDIA Jetson AGX Xavier platform and a D-Link DCS-4625 fisheye camera. The camera was connected directly to the Jetson system via Ethernet through the RTSP protocol. Both components are powered by the vehicle's batteries and Tensor-RT conversion was performed to maximize the algorithm's efficiency, reducing the power consumption to approximately 10 Watts.

The in-shuttle operator on site has already manually been counting passengers using the operator app; hence, validating the automated passenger counting has been done through comparison with data received from the operator's phone data stream (Fig. 4). In Fig. 6, the manual passenger count is seen to the left, where each person getting on the shuttle is entered as 1s in the data stream. Within the same minute as the operator manually counts the 4 entries, the data stream received from the Jetson increases to a count of 5 passengers (4 passengers and the operator). After comparing all data points received within a stable two-day period of operation, the accuracy of the count was further investigated. Only the times where data from our automated approach showed more than 1 person (other than the driver) in the shuttle were extracted and compared. The timestamp is given in UTC time, meaning the actual time was (Copenhagen summer time) 2 hours ahead. By comparing visually the two data streams in Fig. 4, we could see that patterns indicate a similar count of passengers. Initially, an appropriate algorithm to determine the total number of passengers on board was developed that counted only the increases in passenger numbers. However, the oscillating nature of the automated counting data results in a higher total passenger count, combined with the occasional inaccuracies in the exact number of onboard passengers, which pose an ongoing challenge. A post-processing filtering method would help mitigate the error.

D. QUALITATIVE RESULTS

Based on all of the aforementioned information, the main findings of this study suggest that our approach can be

FIGURE 6. Installation in a NAVYA autonomous minibus. Hardware setup consists of a NVIDIA Jetson AGX Xavier device (highlighted in the red box) and a D-Link DCS-4625 fisheye camera (highlighted in blue box).

FIGURE 7. Illustration of a false positive: The green bounding box represents a correctly identified individual, whereas the red bounding box indicates a false positive detection by the algorithm, misidentifying a piece of cloth as a person. This example highlights the challenges in discriminating between actual human figures and objects with similar form factors in complex visual scenes.

successfully implemented across both simple and challenging tasks, while simultaneously maintaining high computational efficiency. Furthermore, it was found that the network's performance is enhanced when the input image resolution is increased to 1,024 × 1.024, at the expense, however, of doubled inference time. Sample results can also be seen for the three datasets in Fig. 9, demonstrating nearly flawless detection across various scenarios, including diverse body poses, orientations, and backgrounds.

However, certain scenarios, such as images of people on a projection screen, low-light conditions, and hard shadows, continue to pose challenges. Our study encounters the challenge of false positives as Fig. 7 illustrates. The algorithm, while adept at identifying individuals with a high degree of accuracy, as denoted by the green bounding box, also exhibits some erroneous classifications. An indicative example is the detection of a non-human object – specifically, a piece of cloth – as a person, highlighted by the red bounding box. Such false positives are not only statistical outliers but also highlight the complexities that these algorithms must navigate. The discriminative power of the algorithm could be tuned to differentiate between human figures and objects similar in shape or size to mitigate the incidence of false positives and enhance the robustness of the detection system in diverse operational environments. Similarly, Fig. 8 illustrates a sequence of extreme and rapid lighting variations caused by shadows and the vehicle's motion. This is a perfect example of a delayed exposure adaptation from the camera sensor, despite featuring WDR. The sudden change in slide 3 results in blown highlights in the image, blending the person's appearance with the vehicle color. This loss of detail results in a failure in the detection of the passenger. Finally, Fig. 9 illustrates a comparison between the proposed method (left side) and [33] (right side) in a crowded scenario. The proposed method correctly detects the two passengers that are partially occluded.

E. ABLATION STUDY

In this section, we conduct ablation experiments to highlight the contribution of each loss function to the overall model performance. As a baseline, we use the second-best model from the comparison table (Table 2) by Duan et al. [33] (first two rows of the Table 3). The second row of the table shows the results of a fine-tuned version of [33] on the MW-R dataset, as in their original paper. Angle, Scale, and the combined Angle + Scale are the results of our implementation in the same dataset.

TABLE 3. Ablation study of various loss functions performance on our dataset.

Loss function	AP_{50}	FPS
Periodic	84.2	27.6
Periodic (finetuned)	87.1	27.5
Angle	91.2	29.2
Scale	89.7	29.3
Angle + Scale	92.3	29.1

VI. CHALLENGES AND LIMITATIONS

The proposed rotation-aware detection framework offers a significant advancement in the field of overhead fisheye passenger detection. However, it's important to acknowledge potential challenges and limitations. The specialized techniques involved in addressing the unique distortions of fisheye images could increase complexity for developers and end-users who are less familiar with advanced computer vision methods. Moreover, as with many object detection approaches, false positives remain a concern, requiring further refinement of the algorithm's ability to discriminate between passengers and other objects. Furthermore, the trade-off between image resolution for better accuracy and

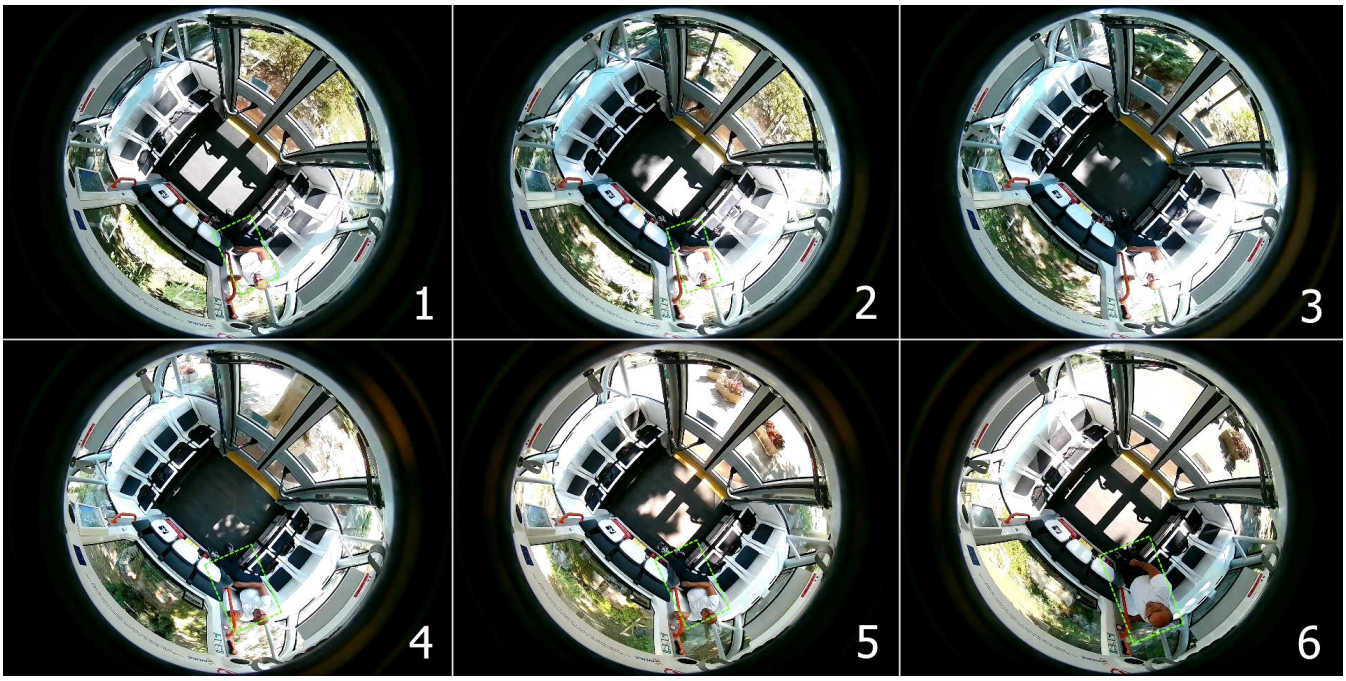


FIGURE 8. Light variations can cause loss of detail in the camera stream, especially without WDR capable sensors.

FIGURE 9. Results from the proposed method (left side) and [33] (right side). The proposed method detects correctly the two passengers that are partially occluded.

the resulting impact on computational efficiency highlights the need to find a practical balance for real-world applications, especially those needing real-time processing. Finally, using a fixed confidence threshold across all datasets might limit adaptability; exploring dynamic thresholding strategies could provide greater flexibility. To enhance the practical deployment of our solution, future research could focus on simplifying implementation for non-expert users, refining its discriminative abilities to minimize false positives in complex environments, exploring computational and energy efficiency optimizations for real-time AV applications, and investigating adaptive thresholding strategies to increase robustness under varying conditions.

VII. CONCLUSION

In this paper, we present a novel algorithm for passenger detection in overhead fisheye images, addressing the limitations of traditional approaches that rely on radially-aligned bounding boxes. Our end-to-end, rotation-aware detection framework utilizes arbitrarily-oriented bounding boxes, providing greater flexibility and enhanced detection accuracy. We introduce a fully convolutional neural network (CNN)

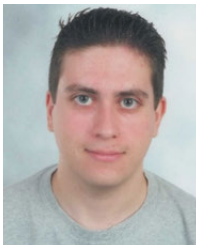
that directly regresses the orientation of each bounding box, combined with a specialized scale and angle loss function. Extensive pre- and post-processing steps are eliminated, reducing computational complexity. Additionally, we present a new dataset designed for in-cabin passenger detection and counting. Our experimental results demonstrate a significant 5.3% improvement in average precision over existing overhead people detection methods, and we validate our approach through real-world deployments in Copenhagen and Geneva, underscoring its value for public transport operators.

REFERENCES

[1] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.
[2] D. T. Nguyen, W. Li, and P. O. Ogunbona, "Human detection from images and videos: A survey," *Pattern Recognit.*, vol. 51, pp. 148–175, Mar. 2016.
[3] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, Jul. 2018.
[4] H. Rashed, E. Mohamed, G. Sistu, V. R. Kumar, C. Eising, A. El-Sallab, and S. Yogamani, "Generalized object detection on fisheye cameras for autonomous driving: Dataset, representations and baseline," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2271–2279.
[5] J. Kang, S. Tariq, H. Oh, and S. S. Woo, "A survey of deep learning-based object detection methods and datasets for overhead imagery," *IEEE Access*, vol. 10, pp. 20118–20134, 2022.
[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.
[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[9] S. Li, M. O. Tezcan, P. Ishwar, and J. Konrad, "Supervised people counting using an overhead fisheye camera," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.
[10] A.-T. Chiang and Y. Wang, "Human detection in fish-eye images using HOG-based detectors over rotated windows," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2014, pp. 1–6.
[11] T. Wang, C.-W. Chang, and Y.-S. Wu, "Template-based people detection using a single downward-viewing fisheye camera," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Nov. 2017, pp. 719–723.
[12] O. Krams and N. Kiryati, "People detection in top-view fisheye imaging," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.
[13] M. Tamura, S. Horiguchi, and T. Murakami, "Omnidirectional pedestrian detection by rotation invariant training," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1989–1998.
[14] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Oriented response networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4961–4970.
[15] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[16] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[17] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1137–1149.
[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
[20] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.
[21] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.
[22] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2Det: A single-shot object detector based on multi-level feature pyramid network," in *Proc. AAAI Conf. Art. Intel. (AAAI)*, vol. 33, Jan. 2019, pp. 9259–9266.
[23] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.
[24] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.
[25] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9756–9765.
[26] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.
[27] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2844–2853.
[28] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3163–3171.
[29] W. Qian, X. Yang, S. Peng, J. Yan, and Y. Guo, "Learning modulated loss for rotated object detection," in *Proc. AAAI Conf. Arti. Intell.*, May 2021, vol. 35, no. 3, pp. 2458–2466.
[30] R. Nosaka, H. Ujiie, and T. Kurokawa, "Orientation-aware regression for oriented bounding box estimation," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.
[31] M. Saito, K. Kitaguchi, G. Kimura, and M. Hashimoto, "People detection and tracking from fish-eye image based on probabilistic appearance model," in *Proc. SICE Annu. Conf.*, Sep. 2011, pp. 435–440.
[32] R. Seidel, A. Apitzsch, and G. Hirtz, "Improved person detection on omnidirectional images with non-maxima suppression," 2018, *arXiv:1805.08503*.
[33] Z. Duan, M. O. Tezcan, H. Nakamura, P. Ishwar, and J. Konrad, "RAPiD: Rotation-aware people detection in overhead fisheye images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2700–2709.
[34] C.-Y. Wang, A. Bochkovskiy, and H. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13024–13033.
[35] N. Ma et al. (Mar. 2018). *Mirror Worlds Challenge*. [Online]. Available: https://www2.icat.vt.edu/mirrorworlds/challenge/index.html
[36] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis.*, Zürich, Switzerland. Cham, Switzerland: Springer, Sep. 2014, pp. 740–755.
[37] A. Haque, A. Alahi, and L. Fei-Fei, "Recurrent attention models for depth-based person identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1229–1238.
[38] D. Tsiktsiris, N. Dimitriou, A. Lalas, M. Dasygenis, K. Votis, and D. Tzovaras, "Real-time abnormal event detection for enhanced security in autonomous shuttles mobility infrastructures," *Sensors*, vol. 20, no. 17, p. 4943, Sep. 2020.
[39] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.
[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "AlexNet," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–9.
[42] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.
[43] A. Mishra, J. Kim, D. Kim, J. Cha, and S. Kim, "An intelligent in-cabin monitoring system in fully autonomous vehicles," in *Proc. Int. SoC Design Conf. (ISOCC)*, Oct. 2020, pp. 61–62.

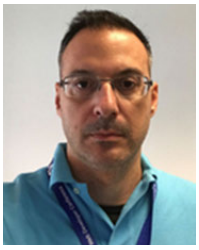
[44] Y.-S. Poon, C.-C. Lin, Y.-H. Liu, and C.-P. Fan, "YOLO-based deep learning design for in-cabin monitoring system with fisheye-lens camera," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2022, pp. 1–4.
[45] A. Mishra, S. Lee, D. Kim, and S. Kim, "In-cabin monitoring system for autonomous vehicles," *Sensors*, vol. 22, no. 12, p. 4360, Jun. 2022.
[46] A. Mishra, J. Cha, and S. Kim, "HCI based in-cabin monitoring system for irregular situations with occupants facial anonymization," in *Proc. 12th Int. Conf. Intell. Human Comput. Interact.* Cham, Switzerland: Springer, Nov. 2020, pp. 380–390.

DIMITRIS TSIKTSIRIS received the Diploma degree in informatics and telecommunications engineering from the Faculty of Engineering, University of Western Macedonia (UOWM), in 2017. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering (ECE), UOWM. He has been a Research Assistant with the Informatics and Technology Institute, Centre for Research and Technology Hellas (ITI/CERTH), since September 2019. His research interests include acceleration on low-powered embedded systems, computer vision, and deep learning approaches.



ANTONIOS LALAS received the Ph.D. degree in electrical and computer engineering, in 2012. From 2012 to 2018, he was an Adjunct Lecturer with the Department of Informatics and Telecommunications Engineering, University of Western Macedonia (UOWM). He was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, AUTH, from 2013 to 2015. From 2020 to 2021, he was an Adjunct Lecturer with the Department of Electrical and Computer Engineering (ECE), UOWM. He is currently a Postdoctoral Researcher with the Information Technologies Institute/Centre for Research and Technology Hellas (ITI/CERTH). His research interests include 5G/6G networks, V2X communications, artificial intelligence, deep neural networks, reconfigurable intelligent surfaces, neuromorphic computing, wireless power transfer, metamaterials, sensor fusion, computational electromagnetics, acoustics, computational fluid dynamics, visualization of physical information, the IoT in relation to autonomous vehicles, counter-UAV, security, cybersecurity, and eHealth domains.

MINAS DASYGENIS received the Ph.D. degree in electrical and computer engineering, in 2005. He is currently an Assistant Professor with the Polytechnic School of Kozani, Department of Electrical and Computer Engineering, University of Western Macedonia, Greece, with a focus on designing embedded systems and accelerators in homogeneous or heterogeneous architectures. He carries over 16 years of teaching experience in operating systems, computer architecture, embedded systems, parallel and distributed systems, and computer networks. He is a Systems' Architect of embedded systems and ICT. He has published more than 85 papers in international journals and conferences and authored three books. He has been a Principal Researcher in three European research projects. His research interests include computer architecture, robotics, embedded and cyber-physical systems, gamification, the Internet of Things, and security and hardware & software cosynthesis. He has been serving as a program committee member or a reviewer for various flagship conferences of embedded systems.



KONSTANTINOS VOTIS received the Ph.D. degree in computer engineering and informatics, in 2011. He has been a Visiting Professor with the Institute for the Future, University of Nicosia, with a focus on blockchain and AI technologies, since October 2019. He is a Senior Researcher (Grade B) at the Information Technologies Institute/Centre for Research and Technologies Hellas and the Director of the Visual Analytics Laboratory. His research interests include human–computer interaction (HCI), information visualization and management of big data, knowledge engineering and decision support systems, the Internet of Things, cybersecurity, and pervasive computing, with major application areas, such as mHealth, eHealth, and personalized healthcare.