IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 April 2024, accepted 4 May 2024, date of publication 13 May 2024, date of current version 22 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3399841

APPLIED RESEARCH

Enhancing the Safety of Autonomous Vehicles: Semi-Supervised Anomaly Detection With Overhead Fisheye Perspective

DIMITRIS TSIKTSIRIS[1,2], ANTONIOS LALAS[1],
MINAS DASYGENIS[2], (Member, IEEE), 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 Autonomous vehicles (AVs) have the potential to revolutionize transportation. However, ensuring passenger safety within these vehicles in the absence of a dedicated onboard authority figure necessitates the development of intelligent, autonomous surveillance systems. This paper presents a novel semi-supervised anomaly detection system specifically designed to enhance safety within autonomous shuttles. Our approach leverages overhead fisheye cameras to provide comprehensive, occlusion-resistant monitoring of the cabin interior. This unique perspective maximizes visibility, even in crowded conditions. Our spatiotemporal autoencoder architecture, composed of both convolutional and reccurrent layers, is trained on extensive unlabeled video data to learn representations of regular passenger behavior using a Center-Weighted Loss (CWL) function that focuses in the cabin's central region, where critical events are most likely to occur. This reduces the potential for false positives triggered by rapid changes on the periphery due to the vehicle's movement. To enhance the system's ability to discriminate between specific safety and security incidents, we introduce a classifier fine-tuned on a labeled subset of our dataset. We evaluate our method's performance through experimentation on a real-world dataset (CERTH-AV) collected with an overhead fisheye camera. Our method demonstrates superior anomaly detection capabilities, achieving the highest Area Under the Curve (AUC) performance on the CERTH-AV dataset. Further comparative evaluations on established benchmarks, including UCF-Crime and ShanghaiTech, validate our system's robustness and adaptability. Finally, we have successfully integrated our method into autonomous minibuses using NVIDIA Jetson embedded systems for real-time processing, demonstrating the practical efficacy of our approach in safeguarding passengers within autonomous vehicles.

INDEX TERMS Anomaly detection, artificial intelligence (AI), autonomous vehicles (AVs), computer vision, edge computing, overhead fisheye imaging.

I. INTRODUCTION

A Autonomous vehicles (AVs) hold the potential to significantly transform transportation, offering increased safety, improved efficiency, and greater accessibility [1]. The autonomous shuttles are operating without a human driver, promising convenient on-demand transportation solutions. However, the absence of an onboard authority figure dedicated to passenger well-being poses unique safety and security risks [2]. The development of systems capable of autonomously monitoring the shuttle's interior and detecting potential threats in real time is crucial to protecting passengers and enabling timely interventions when needed. Overhead fisheye cameras offer a compelling solution for

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

comprehensive interior surveillance in autonomous shuttles [3]. Their wide field of view inherently minimizes occlusions frequently encountered in confined spaces and under dynamic conditions, offering enhanced visibility of the passenger area.

The integration of advanced computer vision and machine learning techniques facilitates the automated detection of anomalous events based on video data, contributing to onboard safety. Autoencoders [4], are employed for anomaly detection by learning compressed representations of normal passenger behavior from video data. Deviations from this learned normality result in higher reconstruction errors, indicating potential anomalies.

While autoencoders excel at detecting unseen anomalies [5], the lack of readily available labeled datasets for anomalous events in AVs presents a challenge. Semi-supervised learning paradigms address this limitation by effectively capitalizing on abundant unlabeled data alongside a smaller set of labeled examples, leading to improved model generalization and robustness. This approach is particularly suitable for autonomous shuttles as collecting large amounts of normal operational data is comparatively simple, while incidents requiring intervention occur less frequently.

Semi-supervised learning, leverages limited or imprecise training labels to learn the underlying patterns in the data, making it particularly suitable for scenarios where obtaining exhaustive and precise annotations is challenging or infeasible, such as in the detection of rare abnormal events [6]. This approach is advantageous in imbalanced datasets because it alleviates the need for extensive labeled data for each class, thus mitigating the bias towards the majority class that is often observed in fully supervised methods. Furthermore, semi-supervised methods are designed to exploit the structure and distribution of the data, enabling them to learn from both labeled and unlabeled data, thereby enhancing their ability to generalize from limited examples of the minority class (abnormal events) [7].

In this work, we propose a novel semi-supervised anomaly detection system specifically designed to enhance safety within the shuttles of autonomous minibuses. These minibuses have a maximum capacity of 10 people, and the seating arrangement typically places passengers within close proximity to the center of the cabin. Our system utilizes overhead fisheye cameras (Figure 1b) for comprehensive cabin monitoring as focusing on the central region encompasses the majority of passenger interactions and potential incident locations and leverages an autoencoder-based framework for anomaly detection. Key contributions of our work include:

- Occlusion-resistant monitoring based on a single fisheye camera perspective, ensuring improved performance under crowded conditions, seamless installation and reduced power consumption.
- Semi-supervised spatiotemporal autoencoder architecture with a center-weighted loss function that learns from “regular” unlabeled video data to detect deviations in passenger behaviour.

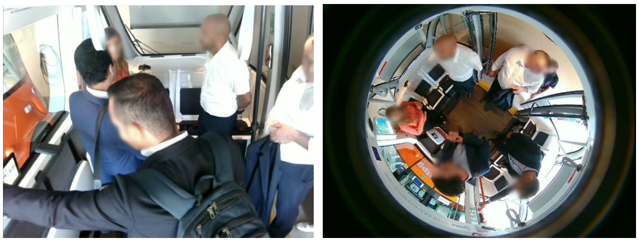

(a) Side perspective (b) Overhead fisheye perspective

FIGURE 1. Examples from different camera setup in an autonomous minibus: (a) Side perspective with a limited field of view and occlusion issues, (b) Overhead fisheye (top-down) camera perspective with a panoramic overview of the cabin.

- Hybrid approach with a head classifier fine-tuned on a smaller labeled subset to identify the abnormal events.
- An extensive evaluation on a real-world dataset, showcasing our method’s ability in detecting critical safety events.

Furthermore, we describe the challenges posed by existing anomaly detection methodologies in the context of fisheye imagery within AVs, such as geometric correction preprocessing and the adaptation of traditional deep learning architectures. By addressing these challenges, our research introduces a fisheye-aware spatiotemporal autoencoder designed for direct application to fisheye image sequences, facilitating enhanced anomaly detection without the need for geometric correction.

The remainder of this work is structured as follows. In Section II, other recent studies in the field of anomaly detection on video surveillance are presented, specifically proposed for overcoming the inherent challenges of weakly supervised settings. Section III offers a detailed overview of the proposed methodology, as well as the training stages that were followed during this work, while Section IV presents the experimental results of this study, along with the system’s real-world implementation in minibuses. Finally, Section V provides a comprehensive discussion on the selection of fisheye cameras, particularly focusing on distortion issues and the justification for prioritizing the central cabin area, while Section VI summarizes the conclusions of this work.

II. RELATED WORK

The integration of deep learning and computer vision for abnormal event detection, particularly in identifying specific actions such as fighting and aggression, has seen notable advancements. Convolutional neural networks (CNNs), specifically, have gained a growing popularity thanks to their high performance in a variety of computer vision applications [8], [9], [10], hence making them also an obvious choice for vision-based human activity recognition tasks.

In the field of weakly supervised learning for anomaly detection in video surveillance, there has been a significant paradigm shift based on the need to alleviate the labor-intensive process of frame-by-frame video annotation [11]. The common thread across this research area is mainly focused around efforts to control video-level annotations, leveraging the inherent spatial and temporal dynamics captured in surveillance footage to determine anomalies with minimal human intervention. This section summarizes the approaches of recent studies in this domain, highlighting the diverse strategies employed to address the challenges associated with minimal manual annotation.

In this context, a key advancement was introduced by Sultani et al. [12] that presented the application of multiple instance learning (MIL) frameworks, establishing the use of video-level annotations for localizing anomalies within videos without precise frame-level labels. This approach laid the key foundations for further exploration of MIL and its capacity to detect anomalous events by exploiting aggregated video observations rather than relying on the granular detail of individual frames.

Graph convolutional networks (GCN) are another emerging advancement in this field that was introduced to refine the approach to handling weakly labeled datasets. Studies in this field [13], [14], [15], [16] have showcased the effectiveness of GCNs in cleaning noisy labels and enhancing the feature extraction process, thereby improving the robustness of anomaly detection models. The key innovation lies in the ability of GCNs to capture complex relationships within the data, enabling a more refined understanding of anomalous patterns.

Moreover, the integration of motion and appearance features has been a key focus for advancing the accuracy of detection models. In this field, existing approaches demonstrate a collective move towards two-stream networks, combining RGB and flow streams [17], [18] to better capture the subtle differences of anomalous behavior, underscoring the recognition that anomalies in video surveillance are often characterized by both unusual appearances and movements, necessitating models that can efficiently navigate this duality.

Adaptive video compression techniques [19], [20], represent another important step towards enhancing real-time detection capabilities. By prioritizing significant events during the preprocessing of surveillance footage, these methods aim to streamline the analysis process, allowing deep learning models to focus on potentially anomalous activities with greater efficiency.

On the other hand, feature amplification mechanisms, introduced by Chen et al. [13], further demonstrate the ongoing efforts to refine anomaly detection accuracy. By amplifying discriminative features and employing magnitude contrastive loss, these models try to overcome the limitations of conventional approaches that rely heavily on feature magnitude for anomaly identification, thereby facilitating a more insightful analysis of surveillance footage, enabling the detection of subtler anomalous activities that might otherwise avoid detection.

Finally, the integration of temporal granularity and spatial features is another approach that has been explored [21], [22], [23] that focuses on anomaly context-dependency, proposing frameworks that address the multifaceted nature of anomalous events. Through the combined learning of motion and appearance features and the application of multiple ranking measures, these studies contribute to a more holistic understanding of what constitutes an anomaly within the vast and varied background of video surveillance data.

The collective effort of the aforementioned research reveals a dynamic and evolving field, with a steady commitment to overcoming the challenges of anomaly detection in weakly supervised settings. However, as previously explained, none of these approaches leverage overhead fisheye cameras for occlusion-resistant monitoring, as opposed to the proposed method. While existing works have explored various methods for anomaly detection in video surveillance, including weakly supervised learning, two-stream networks, and adaptive video compression techniques, they often lack the specific considerations necessary for autonomous vehicles.

Our proposed approach distinguishes itself by leveraging the unique advantages of overhead fisheye cameras for comprehensive cabin monitoring, minimizing occlusion and maximizing visibility within the confined space of an AV. Additionally, the introduction of the Center-Weighted Loss (CWL) function prioritizes the central region of the cabin, where critical events are more likely to occur in the context of autonomous minibuses, further reducing false positives and enhancing the system's effectiveness. The hybrid approach, combining a spatiotemporal autoencoder with a classifier fine-tuned on a labeled subset of data, allows for accurate anomaly detection and distinction between specific types of abnormal events. Furthermore, our system is designed for real-world deployment on resource-constrained embedded platforms, taking into account power consumption and real-time processing requirements.

III. METHODOLOGY

The convolutional autoencoder (CAE) is a specialized form of neural network designed to process data in multiple dimensions, such as images or video frames. It consists of two main parts: an encoder, which compresses the input data into a lower-dimensional latent space; and a decoder, which reconstructs the data from the latent space back to the original input space. The CAE learns to capture the important features of the input data while removing noise or redundancy.

The encoder part of the CAE consists of a series of convolutional layers that apply learned filters to the input data. Convolutional layers are particularly effective for extracting spatial features from images due to their ability to learn hierarchical patterns. The encoder's architecture can be formalized as a function:

$$h_{enc}(x) = f(W_{enc} * x + b_{enc}) \tag{1}$$

where x is the input data, W_{enc} represents the weights of the convolutional filters, b_{enc} denotes the biases, $*$ indicates the convolution operation, and f is a non-linear activation function such as ReLU or Sigmoid.

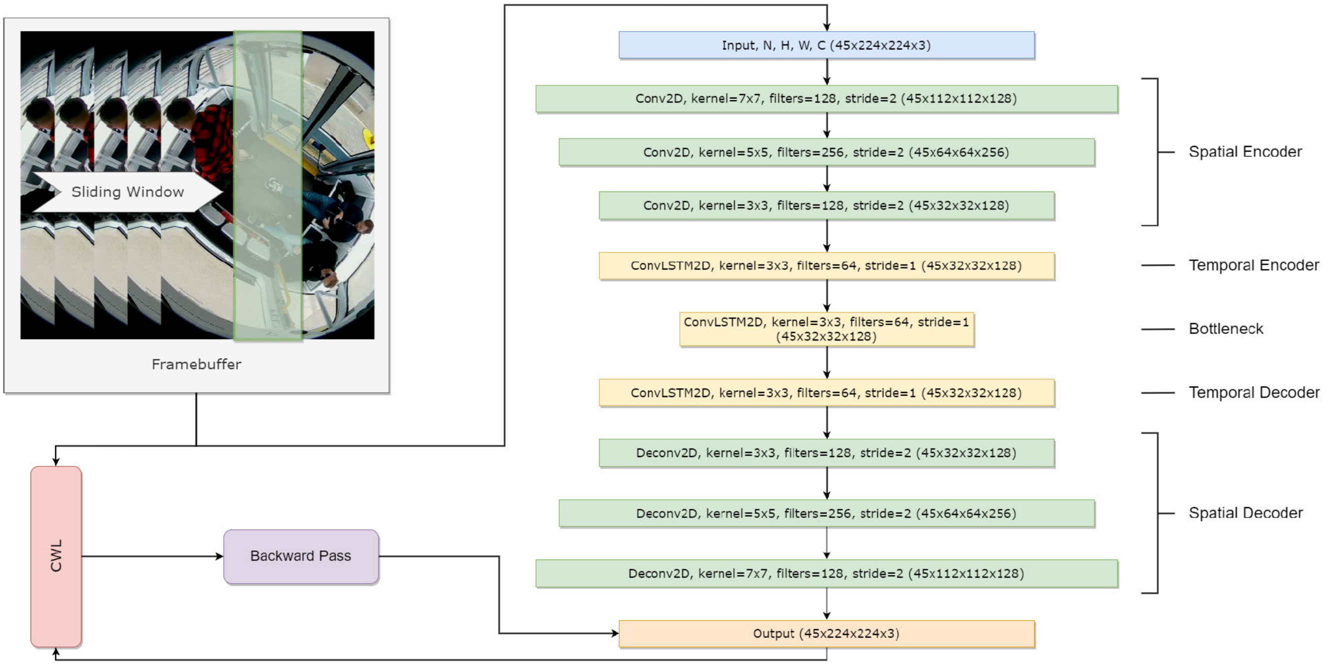

Sliding Window
Framebuffer
Input, N, H, W, C (45x224x224x3)
Conv2D, kernel=7x7, filters=128, stride=2 (45x112x112x128)
Conv2D, kernel=5x5, filters=256, stride=2 (45x64x64x256)
Conv2D, kernel=3x3, filters=128, stride=2 (45x32x32x128)
ConvLSTM2D, kernel=3x3, filters=64, stride=1 (45x32x32x128)
ConvLSTM2D, kernel=3x3, filters=64, stride=1 (45x32x32x128)
ConvLSTM2D, kernel=3x3, filters=64, stride=1 (45x32x32x128)
Deconv2D, kernel=3x3, filters=128, stride=2 (45x32x32x128)
Deconv2D, kernel=5x5, filters=256, stride=2 (45x64x64x256)
Deconv2D, kernel=7x7, filters=128, stride=2 (45x112x112x128)
Output (45x224x224x3)
Spatial Encoder
Temporal Encoder
Bottleneck
Temporal Decoder
Spatial Decoder
CWL
Backward Pass

FIGURE 2. Model architecture of the autoencoder: The convolutional layers are spatial encoders, followed by temporal encoder and decoder. Bottleneck compress the features to eliminate non useful information. At the end, we perform spatial decoding, reconstructing the input image to the same format.

The ConvLSTM2D layer is a recurrent layer that processes data in both space and time. It is specifically designed for problems where the context in both dimensions is crucial, such as videos. The layer not only applies a convolution operation to the input data but also maintains a hidden state that captures temporal information. The ConvLSTM2D layer can be expressed mathematically as:

$$h_t, c_t = \mathrm{ConvLSTM2D}(h_{t-1}, c_{t-1}, x_t) \tag{2}$$

where h_t is the hidden state at time t, c_t is the cell state at time t, and x_t is the input at time t.

The decoder mirrors the encoder structure but uses deconvolutional (transposed convolution) layers to reconstruct the input data from the latent space representation. The reconstruction can be quantified using the proposed CWL loss.

A. CENTER-WEIGHTED LOSS

Regarding our loss function, we combined the traditional mean squared error (MSE) loss with a spatial weighting mechanism that assigns higher weights to pixels closer to the center of the image. This weighting can help the autoencoder focus more on accurately reconstructing the central part of the image, which has a higher impact based on the distribution of passengers in the cabin space. The proposed loss function is defined as follows:

$$\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W} w(i,j)\cdot(\mathbf{Y}_{ij} - \hat{\mathbf{Y}}_{ij})^2,$$

where $\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}})$ is the loss function, $\mathbf{Y}$ is the ground truth image, $\hat{\mathbf{Y}}$ is the reconstructed image produced by the autoencoder, and H and W are the height and width of the images, respectively. The weight function $w(i,j)$ is defined using a Gaussian distribution:

$$w(i,j) = \exp\left(-\frac{(i-i_c)^2 + (j-j_c)^2}{2\sigma^2}\right),$$

with (i_c, j_c) representing the coordinates of the center of the image and σ controlling the spread of the Gaussian function. The normalization factor N, often the total number of pixels in the image, is used to keep the loss value scale consistent.

Incorporating the Gaussian function into the weighting mechanism allows for a smooth transition of importance from the center towards the edges of the image, which aligns with the goal of enhancing focus on the central areas during the autoencoder's training process.

B. TRAINING STAGES

At the first stage (Figure 2), the CAE is trained on regular 'normal' events using back-propagation to minimize the reconstruction loss. The process optimizes the weights and biases to capture the regular patterns of the input data. After the initial training, a second stage of supervised training follows with a slightly modified architecture (Figure 3). In this training stage, we use the encoder part of the model along with a classification head for supervised training. We perform fine tuning via transfer learning using the weights from the previous training session, which can ensure the ability of the encoder to compress critical features into the latent space to improve the robustness of our method.

The head is trained using categorical cross-entropy loss:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{3}$$

where M is the number of classes, $y_{o,c}$ indicates the presence of class c in observation o, and $p_{o,c}$ is the predicted probability of class c for observation o.

The two-phase training strategy of the CAE allows the model to be capable of distinguishing irregular changes in the sequences and then to identify deviations from this baseline as anomalies. The integration of ConvLSTM2D layers enables the model to capture temporal dependencies in addition to the spatial features learned by the Conv2D layers. This method provides a robust approach to anomaly detection in video sequences.

IV. RESULTS

In this section, we present the datasets, the evaluation metrics, the parameter settings and the experimental results of our approach. Moreover, we present the result of a real-world deployment of an NVIDIA Jetson embedded system, capable of detecting four types of abnormal events: bag snatching, falling down, fighting and vandalism. The system is installed and operates inside autonomous minibuses and utilizes the proposed method for abnormal event detection.

A. DATASETS

There are various well-established datasets for anomalous activity recognition within video surveillance footage, with the most prominent among them including the UCSD Pedestrian [25], Subway [26] and CUHK Avenue [27] datasets. However, despite their extensive application, these datasets show various limitations, including the simplicity of the depicted scenes, their restricted range of anomalous activities and the lack of detailed spatial annotations, that may lead to unsatisfactory outcomes in real-world scenarios.

On the other hand, datasets such as the UCF-Crime [12] and the ShanghaiTech [24] offer a comprehensive collection of videos sourced from various online platforms, recorded across multiple surveillance systems under a wide array of environmental conditions, thereby introducing additional layers of complexity. More specifically, the UCF-Crime dataset includes approximately 1900 extensive, unedited videos, evenly split between normal and abnormal events, and covers 13 types of real-world anomalies, including but not limited to abuse, burglary and vandalism. Additionally, the ShanghaiTech dataset aims to address real-world applicability issues by including anomalies characterized by sudden movements, like chases and fights, including 130 anomalous events across 13 settings in 437 videos, totaling over 270,000 frames for training. It specifically labels unusual activities such as bag snatching and unauthorized vehicle use.

However, while the UCF-Crime dataset initially contributed video-level annotation, which was especially useful for weakly supervised learning approaches, the ShanghaiTech dataset has been utilized for unsupervised learning to determine regular patterns. The availability of frame-level annotations for both datasets facilitates the adoption of fully supervised learning methods. The UCF-Crime dataset, with its varied activity rate and environmental settings, is particularly practical for anomaly detection tasks, enhancing the development of surveillance systems operating in real-world environments.

Besides the two benchmark datasets mentioned previously, we collected a real-world dataset (CERTH-AV) using the D-Link DCS-4625 fisheye camera with an overhead panoramic perspective. The camera has a dome design, featuring a 5-megapixel 1/2.5'' CMOS sensor, paired with a 1.37 mm F2.0 fisheye lens. It has a maximum image resolution of 2560×1920 pixels and has Wide Dynamic Range (WDR) support along with IR lighting for night vision. During the data collection, several abnormal events were simulated (bagsnatch, falldown, fighting, vandalism) with a duration of approximately 30 minutes. Moreover, we collected regular vehicle operation, whether stationary or in motion, with various lighting conditions and passengers. Table 2 presents some metrics about these datasets. It is important to note that in the initial training phase of the CAE only the regular ''normal'' class is used, while a subset of the class is used for fine-tuning the hybrid approach. Consents were obtained from all passengers who contributed to this dataset for the purpose of this research.

B. EVALUATION METRICS

The experimental results of this work were evaluated against established benchmarks in the field using universally recognized metrics for detecting anomalies. To keep consistency with previous studies in anomaly detection [24], [28], the findings are expressed through the frame-level Area Under the Curve (AUC) to facilitate a comparison of performance levels, with a higher AUC indicating better detection capabilities. Given the challenges posed by datasets where anomalies are rare, AUC is preferred over accuracy as it offers a more refined assessment [29]. Additionally, the model's ability to distinguish between different types of anomalies was assessed using confusion matrices (Figure 4) and the accuracy metric.

C. ABLATION EXPERIMENTS

To investigate the influence of different weighting functions on the anomaly detection performance, we conducted an ablation study comparing the Gaussian distribution with three alternative functions: Linear Decay (Equation 4), Inverse Distance (Equation 5), and Sigmoid (Equation 6).

$$w(i,j) = 1 - \frac{\sqrt{(i-i_c)^2 + (j-j_c)^2}}{d_{max}} \tag{4}$$

$$w(i,j) = \frac{1}{1 + \sqrt{(i-i_c)^2 + (j-j_c)^2}} \tag{5}$$

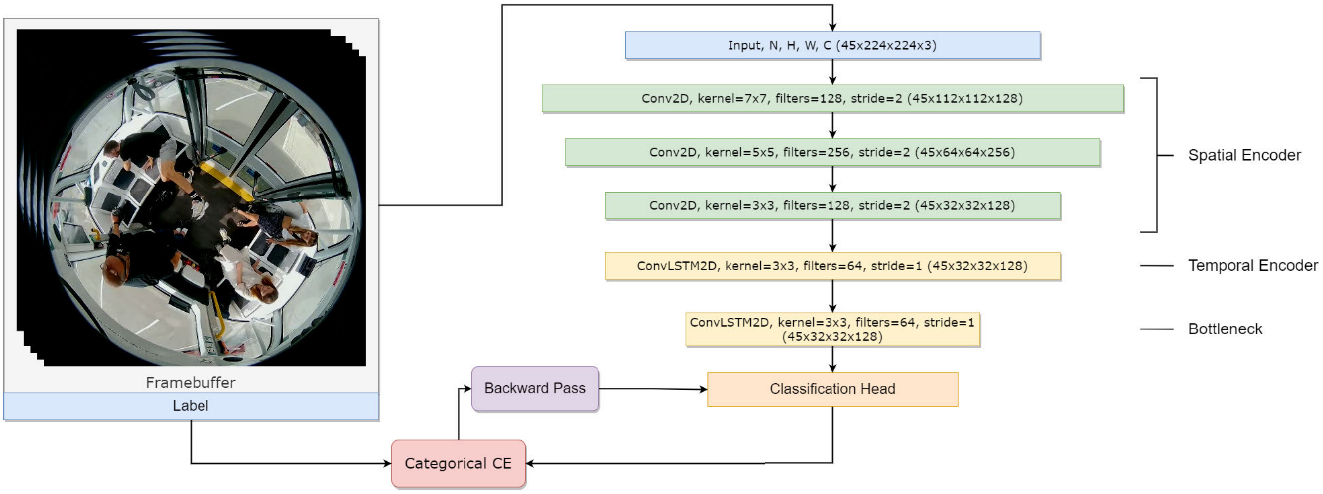

Framebuffer
Label
Input, N, H, W, C (45x224x224x3)
Conv2D, kernel=7x7, filters=128, stride=2 (45x112x112x128)
Conv2D, kernel=5x5, filters=256, stride=2 (45x64x64x256)
Conv2D, kernel=3x3, filters=128, stride=2 (45x32x32x128)
ConvLSTM2D, kernel=3x3, filters=64, stride=1 (45x32x32x128)
ConvLSTM2D, kernel=3x3, filters=64, stride=1 (45x32x32x128)
Spatial Encoder
Temporal Encoder
Bottleneck
Backward Pass
Classification Head
Categorical CE

FIGURE 3. Fine-tuning hybrid classifier: The pretrained encoder weights from the previous stage are transferred and a classification head is added for detecting the abnormal events.

TABLE 1. AUC results evaluation of the proposed method on the datasets of CERTH-AV and benchmark datasets UCF-Crime and shanghai tech.

Work	Method	CERTH-AV	UCF-crime [12]	ShanghaiTech [24]
Sultani *et al.* [12]	DeepMII.	0.622	0.7541	–
Liu *et al.* [11]	MLEP	0.651	–	0.768
Zhong *et al.* [14]	GCN-C3D	0.702	0.810	0.764
Zhong *et al.* [14]	GCN-TSNRGB	0.736	0.821	0.844
Zhong *et al.* [14]	GCN-TSNOptical-Flow	0.718	0.780	0.841
Gianchandani *et al.* [22]	Spatiotemporal	0.672	0.630	–
Hao *et al.* [17]	TSNRGB+Optical-Flow	0.724	0.812	**0.967**
Shreyas *et al.* [19]	AC-MIL	–	0.798	–
Zaheer *et al.* [16]	Binary clustering	0.721	0.782	0.841
Dubey *et al.* [21]	DMRMS	–	0.819	0.685
Majhi *et al.* [23]	Two-level attention	0.741	0.821	–
Ullah *et al.* [20]	CNN-BDLSTM	0.745	0.855	–
Cao *et al.* [15]	WAGCN	0.695	0.846	0.960
Thakare *et al.* [18]	TCC	0.756	0.844	–
Chen *et al.* [13]	MGFN	0.732	**0.869**	**–**
Proposed Method	CAE-Hybrid	**0.878**	0.852	0.927

TABLE 2. Statistics of the overhead fisheye camera dataset (CERTH-AV).

Class	Size (GB)	Samples	Duration	FPS
bagsnatch	4.5	125	0h 30m 40s	15
falldown	3.9	109	0h 26m 55s	15
fighting	5.1	142	0h 35m 35s	15
normal	39.5	1098	4h 15m 20s	5-15
vandalism	6.2	172	0h 40m 35s	15

TABLE 3. AUC performance with different weighting functions.

Weighting Function	AUC
Gaussian Distribution	0.878
Linear Decay	0.862
Inverse Distance	0.851
Sigmoid Function	0.871

$$w(i,j) = \frac{1}{1+e^{-k\left(\sqrt{(i-i_c)^2+(j-j_c)^2}-r\right)}} \tag{6}$$

where w is the weight assigned to the pixel at coordinates (i,j); coordinates (i_c, j_c) represent the center of the image and (d_{max}) is the maximum distance from the center of the image to any pixel (typically the diagonal distance). Moreover, in equation 6, k controls the steepness of the sigmoid function, while r controls the midpoint of the transition between high and low weights.

As shown in Table 3, the Gaussian distribution achieved the highest AUC score, indicating the best overall performance among the evaluated functions. The Sigmoid function demonstrated a close performance, while the Linear Decay and Inverse Distance functions resulted in slightly lower AUC scores.

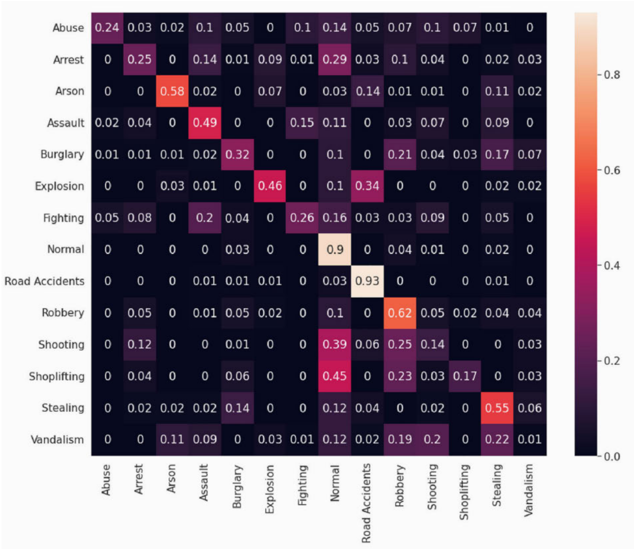


FIGURE 4. Confusion matrix across multiple categories: The diagonal represents accurate predictions and off-diagonal cells indicate false predictions. The intensity of the color corresponds to the normalized frequency of predictions, highlighting the precision and misclassification rates between different classes.

TABLE 4. Precision, recall, and F1-Score for each class tested in UCF-Crime dataset.

Class	Precision	Recall	F1-Score
Abuse	0.7500	0.2449	0.3692
Arrest	0.3906	0.2475	0.3030
Arson	0.7532	0.5859	0.6591
Assault	0.4414	0.4900	0.4645
Burglary	0.4444	0.3232	0.3743
Explosion	0.6765	0.4694	0.5542
Fighting	0.4906	0.2626	0.3421
Normal	0.2961	0.9000	0.4455
Road Accidents	0.5671	0.9300	0.7045
Robbery	0.3483	0.6200	0.4460
Shooting	0.1750	0.1400	0.1556
Shoplifting	0.5862	0.1683	0.2615
Stealing	0.4198	0.5556	0.4783
Vandalism	0.0323	0.0100	0.0153

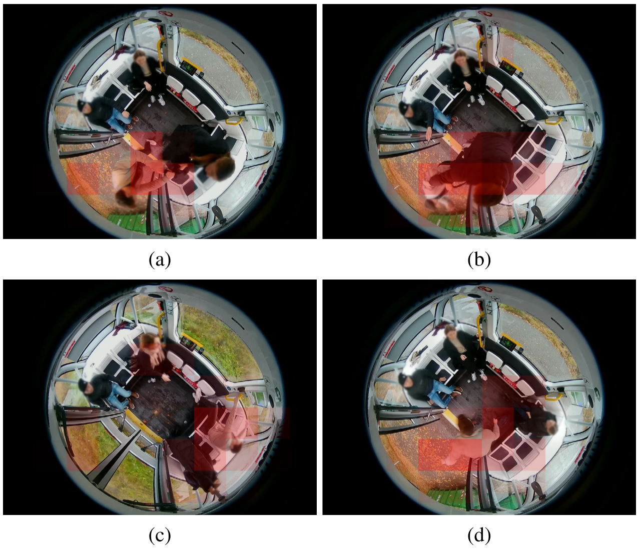

(a)
(b)
(c)
(d)

FIGURE 5. Real-time detection of abnormal events with activation maps visualization in red shades: (a) and (b) indicate a fighting event, (c) passengers falling down due to a sudden deceleration (breaking) of the vehicle, (d) bag snatching (stealing) event.

The results suggest that the Gaussian distribution's smooth transition of weights from the center to the periphery effectively balances the focus on the central region with consideration of the surrounding context. This characteristic is beneficial in capturing anomalies that may involve interactions between passengers or events that start in the center and propagate outward. The Sigmoid function, with its steeper slope, places a stronger emphasis on the central region. This could be advantageous in scenarios where anomalies primarily occur within a well-defined central area. However, it may also lead to overfocusing on events near the edges or interactions across a larger portion of the cabin. Finally, the Linear Decay and Inverse Distance functions, while performing adequately, may not capture the subtle movements of passenger behavior and interactions as effectively as the Gaussian and Sigmoid functions. Their simpler weighting schemes may not fully account for the spatial characteristics of the fisheye image and the distribution of passengers within the cabin.

D. TRAINING SETTINGS

To achieve peak performance, we explored the optimization of critical hyperparameters. We experimented with parameters such as frame-skipping, data normalization, temporal-length, temporal-stride, and the potential benefits of randomized runtime data augmentation for temporal sequence generation. We also carefully examine the potential for enhancements within the deep network architecture itself. We trained our model with a batch size of 16, a temporal-length of 45 frames and an input image dimension of 224×224 with 3 channels. To prevent overfitting, we implement dynamic frame-skipping on input videos with a stride between 1 and 5, and random runtime augmentations are applied on image sequences with a probability of 0.3. Train, validation, and test sets are randomized to ensure a robust and unbiased evaluation.

Due to the large number of video data, the bottleneck of our training process was the data loader. Operations for decompressing videos, preprocessing, image transformations and transferring data from the CPU to the GPU had a significant performance impact on the training process. In order to maximize computational efficiency and streamline the training process, we integrate NVIDIA DALI for accelerated data loading. This optimization ensures that our hardware isn't bottlenecked by data preparation. Training is conducted using the Adam optimizer with a learning rate of 0.0001. To harness the latest in computational power, this training process is powered by a system equipped with the high-end

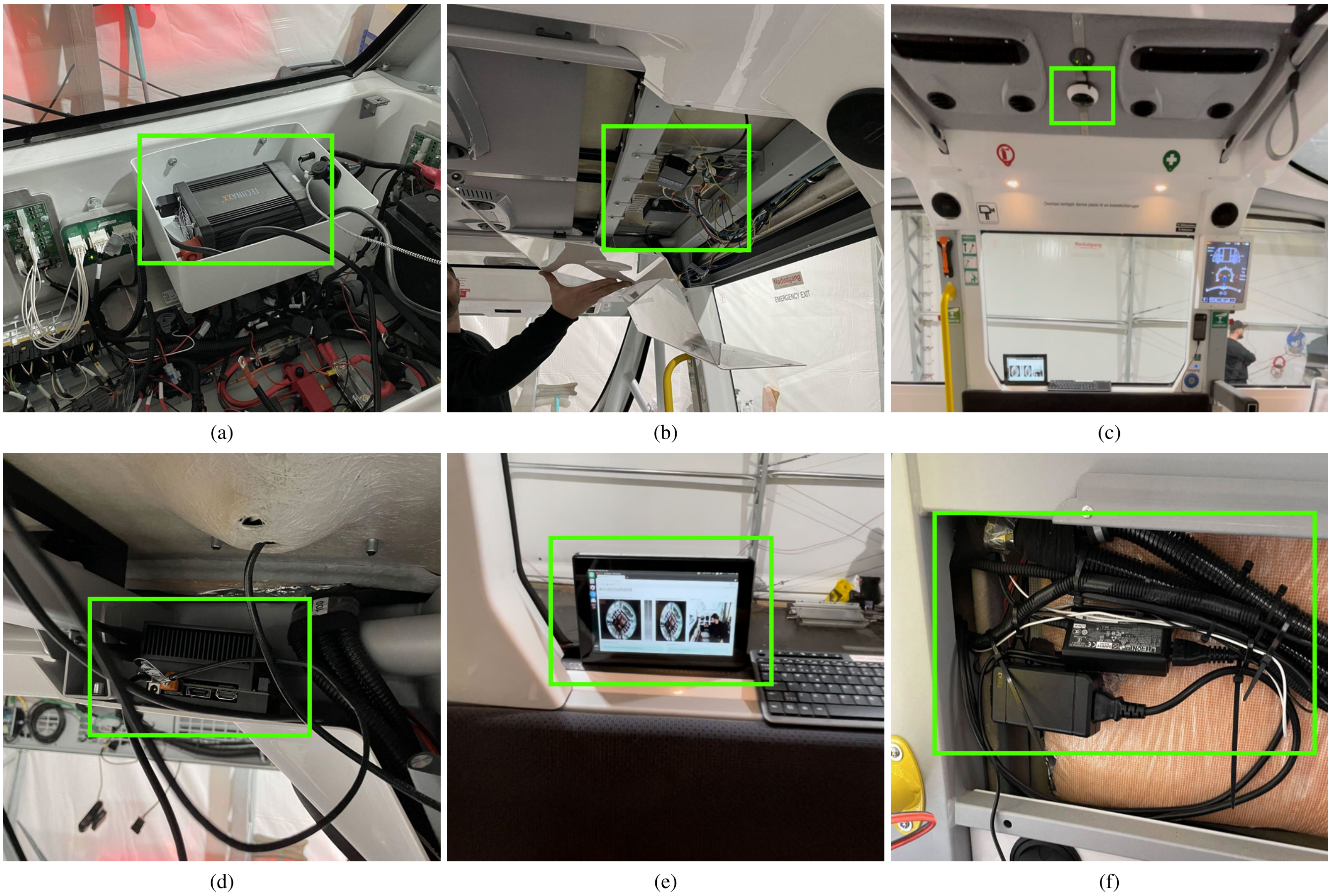


FIGURE 6. Real-world installation on NAVYA autonomous minibuses: (a) Inventer, (b) Installation area - panel, (c) Overhead fisheye camera (D-Link DCS 4625), (d) Location of the NVIDIA Jetson AGX Xavier embedded system, (e) Monitoring screen for passengers, (f) Power-supply cables hidden behind the vehicle's panel.

NVIDIA 4090 GTX 24GB GPU, an Intel Core i7-13700K processor, and 64 GB of RAM.

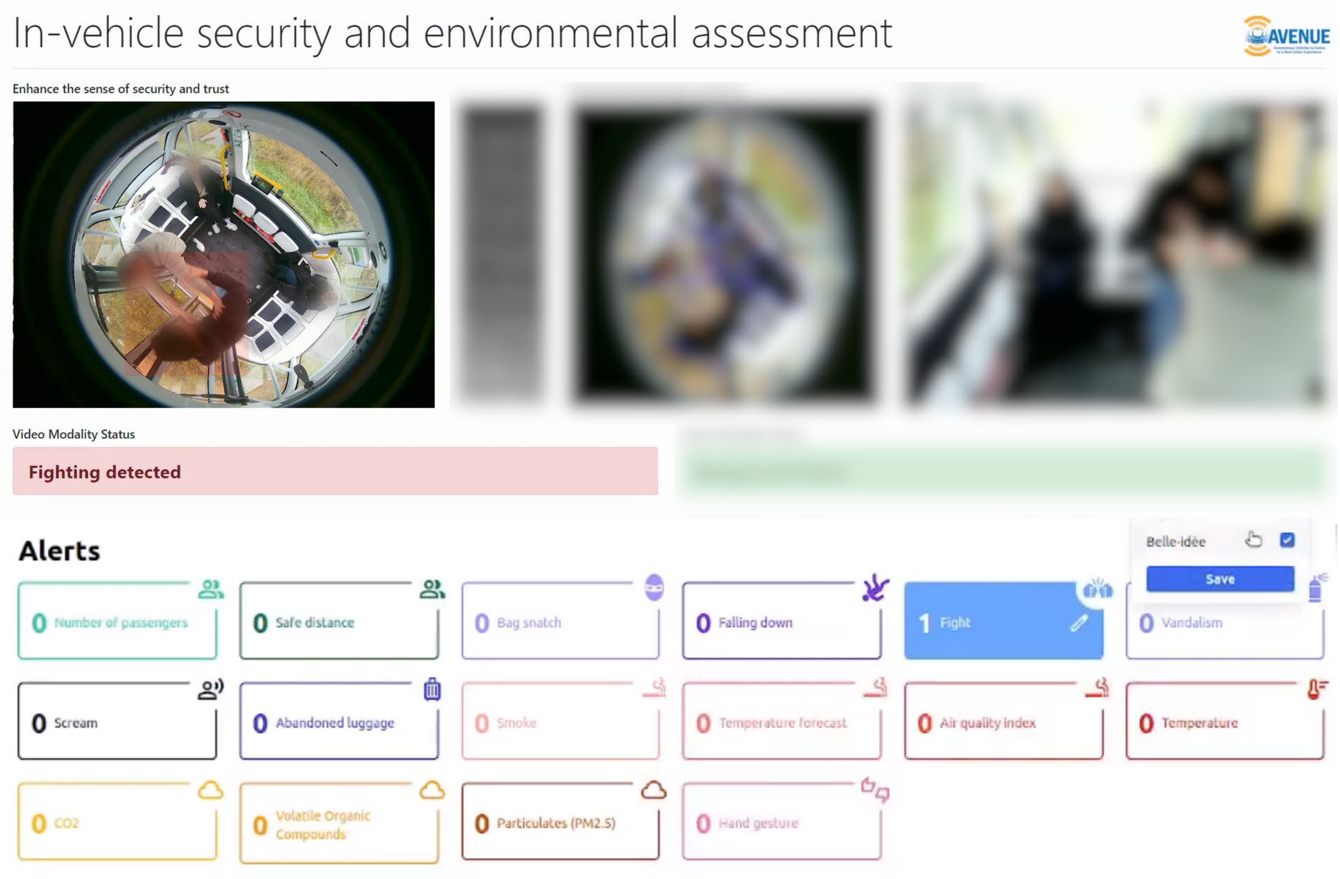
E. MAIN RESULTS

Table 1 presents the comparative evaluation of various anomaly detection methods based on the Area Under the Curve (AUC) performance across three distinct datasets: CERTH-AV, UCF-Crime, and ShanghaiTech. These datasets are benchmarks in the domain of anomaly detection within video surveillance, each presenting unique challenges and complexities. The table illustrates a notable evolution in methodological complexity and specificity, with recent methods like Cao et al. [15] in 2022 WAGCN and Chen et al. [13] in 2023 MGFN showcasing significant strides in performance, particularly on the UCF-Crime dataset. This reflects a trend towards more adaptive anomaly detection mechanisms, capable of handling the diverse and complex scenarios presented by these datasets. Our proposed method, CAE-Hybrid had the highest AUC of 0.878 on CERTH-AV, alongside competitive performances on UCF-Crime and ShanghaiTech. This suggests a robust and versatile approach, capable of overcoming the challenges related to the overhead fisheye camera perspective due to its incorporation of the center-weighted function loss and hybrid approach.

It is important to note that the absence of results for certain methods on our CERTH-AV dataset, is attributed to the inability to reproduce the code for the respective methods. In cases where code was not available, implementations were based on the authors' interpretations of the published methodologies, which might not fully capture the original intent of these approaches. By emphasizing the central aspects of the image more significantly than the peripheral ones, the robustness of this method is increased by reducing false positives and focusing on core features that are indicative of anomalous behavior, improving its sensitivity and specificity in anomaly detection tasks.

F. VALIDATION IN NAVYA AUTONOMOUS VEHICLES

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, as presented in Figures 6c and 6d. 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.



In-vehicle security and environmental assessment

Enhance the sense of security and trust

Video Modality Status

Fighting detected

Alerts

0	Number of passengers
0	Safe distance
0	Bag snatch
0	Falling down
1	Fight
0	Vandalism
0	Scream
0	Abandoned luggage
0	Smoke
0	Temperature forecast
0	Air quality index
0	Temperature
0	CO2
0	Volatile Organic Compounds
0	Particulates (PM2.5)
0	Hand gesture

FIGURE 7. Dashboard showcasing a fighting abnormal event detection using the proposed method.

TABLE 5. Number of samples validated, along with the improved accuracy and F1-Score metrics for each class.

Class	Samples	Accuracy	F1-Score
bagsnatch	15	0.93	0.9310
falldown	12	0.86	0.8800
fighting	13	0.92	0.9240
vandalism	15	0.94	0.9355

The system was installed and validated in real NAVYA autonomous vehicles by Amobility (HOLO) and Transports Publics Genevois (TPG) operators, resulting in the validation results presented in Table 5. As observed by the table, these results indicate that the proposed semi-supervised anomaly detection algorithm is able to identify most of the performed scenarios (bagsnatch, falldown, fighting and vandalism) with an accuracy of approximately 91%.

Figure 5 demonstrates snapshots during the real-time detection of various abnormal events in HOLO vehicle P109, route Slagelse in Copenhagen, Denmark. For visualization purposes, activation maps are overlaid to highlight regions of interest, denoted by red shades, where anomalous activities are detected. Finally, as depicted in Figure 7, the abnormal event identification has also been captured in the operator's dashboard in real-time, raising appropriate alerts for the authorities, thereby enhancing the overall safety and trust of onboard passengers.

G. COMPUTATIONAL EFFICIENCY

The proposed semi-supervised anomaly detection system is designed to operate in real-time within the constraints of an embedded system, ensuring prompt detection of critical events in autonomous shuttles. To optimize computational efficiency, we employ several strategies. Firstly, the convolutional autoencoder architecture utilizes depth-wise separable convolutions, which significantly reduce parameter size compared to standard convolutions. Moreover, network optimizations via tools like TensorRT enable the model to run efficiently on the NVIDIA Jetson platform, which possesses limited computational resources when compared to desktop GPUs. Experimental results demonstrate that our approach achieves a processing speed of approximately 37 frames per second (FPS) using the Performance (MAX-N) mode on the NVIDIA Jetson AGX Xavier embedded system. However, we are restricting the processing framerate to 15 FPS, accumulating a sliding buffer of 3 seconds. The use of TensorRT optimizations (layer fusion, precision reduction, kernel auto-tuning, and dynamic tensor memory optimization) along with a reduced processing frame rate has reduced the power consumption to an acceptable level for the AV's battery. The NVIDIA Jetson AGX Xavier platform, running the anomaly detection algorithms, operates at an average power of approximately 10 watts, as measured by the tegrastats utility. The D-Link DCS-4625 fisheye camera has an average power consumption of 3.5 watts, based on the manufacturer's specifications. Based on our experiments, we consider it sufficient for real-time deployment within power-constrained autonomous shuttles, allowing for timely notifications in case of critical safety or security events.

H. ETHICAL CONSIDERATIONS

While continuous surveillance within autonomous vehicles offers significant potential to improve passenger safety, it's crucial to acknowledge the ethical implications involved. The focus on edge computing within our system mitigates a significant portion of privacy concerns, as no sensitive raw video data is transmitted outside the vehicle. Only the necessary metadata for operator notifications are sent, respecting the privacy of the passengers. However, it is still important to consider techniques that further preserve the anonymity of individuals within the video feed while maintaining robust anomaly detection [30], such as blurring faces or Federated Learning [31]. Additionally, clear policies on data retention and access must be established in alignment with relevant regulations (e.g., GDPR). Transparency with passengers regarding the use of the system and the protection of their personal data is vital to establishing public trust. The development of such systems demands a careful balance between safeguarding passengers and protecting individual privacy rights.

V. DISCUSSION

Our current work focuses specifically on demonstrating the effectiveness of overhead fisheye cameras as a standalone solution for anomaly detection in the context of autonomous minibuses. Due to the relatively confined space, a single overhead fisheye camera can provide sufficient coverage of the entire cabin similar to a multi-camera setup.

However, multiple cameras would have a negative impact on power consumption, due to the additional hardware as well as the increased processing complexity. This is a significant constraint for battery-powered autonomous vehicles.

Considering the practical aspects of implementation, the NAVYA autonomous minibuses used in our study already come equipped with a pre-installed overhead fisheye camera. Therefore, our approach can take advantage of the existing hardware infrastructure, making it a more cost-effective and readily deployable solution.

Comparative evaluations with traditional narrow-field cameras indicate that fisheye cameras capture a broader view with fewer installation points, enhancing the system's efficiency and cost-effectiveness [32].

VI. CONCLUSION

This study has presented a novel semi-supervised anomaly detection system specifically designed to enhance safety within autonomous shuttles. The utilization of overhead fisheye cameras offers a comprehensive and occlusion-resistant perspective on the cabin interior, maximizing visibility and minimizing the potential for limited vision spots. Our spatiotemporal autoencoder architecture, trained on unlabeled video data, effectively models typical passenger behavior, forming a baseline for anomaly detection. The integration of a classifier, fine-tuned with the CWL function, enhances the system's capability to discriminate between specific types of anomalies and regular activity, prioritizing events occurring in the central image region. Extensive evaluations on the real-world CERTH-AV dataset demonstrate the superior performance of our method, as evidenced by its high AUC score. Additionally, the method's strong results on benchmark datasets like UCF-Crime and ShanghaiTech underscore its robustness and adaptability to diverse surveillance scenarios. The successful real-world deployment of our system on NVIDIA Jetson embedded systems within autonomous minibuses validates its applicability and effectiveness in practical settings. Future work includes exploring the fusion of data from multiple sensor modalities, such as depth or thermal cameras, for enhanced anomaly detection. Additionally, investigating the potential of reinforcement learning techniques within the semi-supervised framework could offer further refinement of the anomaly detection process. Finally, this work contributes to the development of intelligent surveillance systems designed to protect passengers and enable the safe, widespread adoption of autonomous shuttles.

REFERENCES

[1] S. A. Bagloee, M. Tavana, M. Asadi, and T. Oliver, "Autonomous vehicles: Challenges, opportunities, and future implications for transportation policies," *J. Mod. Transp.*, vol. 24, no. 4, pp. 284–303, Dec. 2016.

[2] D. J. Fagnant and K. Kockelman, "Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations," *Transp. Res. A, Policy Pract.*, vol. 77, pp. 167–181, Jul. 2015.

[3] V. R. Kumar, C. Eising, C. Witt, and S. K. Yogamani, "Surround-view fisheye camera perception for automated driving: Overview, survey & challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 3638–3659, Apr. 2023.

[4] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal AutoEncoder for video anomaly detection," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1933–1941.

[5] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 665–674.

[6] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft, "Deep semi-supervised anomaly detection," 2019, *arXiv:1906.02694*.

[7] A. Kumar and Y. S. Rawat, "End-to-end semi-supervised learning for video action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14680–14690.

[8] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences," *Int. J. Speech Technol.*, vol. 51, no. 2, pp. 690–712, Feb. 2021.

[9] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz, L. de Paula Veronese, T. Oliveira-Santos, and A. F. De Souza, "Self-driving cars: A survey," *Expert Syst. Appl.*, vol. 165, Mar. 2021, Art. no. 113816.

[10] 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.

[11] W. Liu, W. Luo, Z. Li, P. Zhao, and S. Gao, "Margin learning embedded prediction for video anomaly detection with a few anomalies," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3023–3030.

[12] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[13] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 1, pp. 387–395.

[14] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1237–1246.

[15] C. Cao, X. Zhang, S. Zhang, P. Wang, and Y. Zhang, "Adaptive graph convolutional networks for weakly supervised anomaly detection in videos," *IEEE Signal Process. Lett.*, vol. 29, pp. 2497–2501, 2022.

[16] M. Zaigham Zaheer, J.-H. Lee, M. Astrid, A. Mahmood, and S.-I. Lee, "Cleaning label noise with clusters for minimally supervised anomaly detection," 2021, *arXiv:2104.14770*.

[17] W. Hao, R. Zhang, S. Li, J. Li, F. Li, S. Zhao, and W. Zhang, "Anomaly event detection in security surveillance using two-stream based model," *Secur. Commun. Netw.*, vol. 2020, pp. 1–15, Aug. 2020.

[18] K. V. Thakare, N. Sharma, D. P. Dogra, H. Choi, and I.-J. Kim, "A multi-stream deep neural network with late fuzzy fusion for real-world anomaly detection," *Expert Syst. Appl.*, vol. 201, Sep. 2022, Art. no. 117030.

[19] D. G. Shreyas, S. Raksha, and B. G. Prasad, "Implementation of an anomalous human activity recognition system," *Social Netw. Comput. Sci.*, vol. 1, no. 3, pp. 1–10, May 2020.

[20] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools Appl.*, vol. 80, no. 11, pp. 16979–16995, May 2021.

[21] S. Dubey, A. Boragule, J. Gwak, and M. Jeon, "Anomalous event recognition in videos based on joint learning of motion and appearance with multiple ranking measures," *Appl. Sci.*, vol. 11, no. 3, p. 1344, Feb. 2021.

[22] U. Gianchandani, P. Tirupattur, and M. Shah, "Weakly-supervised spatiotemporal anomaly detection," Univ. Central Florida Center Res. Comput. Vis. REU, Orlando, FL, USA, 2019. [Online]. Available: https://www.crcv.ucf.edu/wp-content/uploads/2019/08/Nsf_Reu2019_Urvi_G_Report.pdf

[23] S. Majhi, S. Das, F. Brémond, R. Dash, and P. K. Sa, "Weakly-supervised joint anomaly detection and classification," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–7.

[24] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 341–349.

[25] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[26] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[27] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720–2727.

[28] B. Wan, W. Jiang, Y. Fang, Z. Luo, and G. Ding, "Anomaly detection in video sequences: A benchmark and computational model," *IET Image Process.*, vol. 15, no. 14, pp. 3454–3465, Dec. 2021.

[29] C. X. Ling, J. Huang, and H. Zhang, "AUC: A better measure than accuracy in comparing learning algorithms," in *Advances in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 2671, Y. Xiang and B. Chaib-Draa, Eds. Berlin, Germany: Springer, 2003, pp. 329–341.

[30] M. Ye, W. Shen, J. Zhang, Y. Yang, and B. Du, "SecureReID: Privacy-preserving anonymization for person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2840–2853, 2024.

[31] M. K. Hasan, N. Jahan, M. Z. A. Nazri, S. Islam, M. A. Khan, A. I. Alzahrani, N. Alalwan, and Y. Nam, "Federated learning for computational offloading and resource management of vehicular edge computing in 6G-V2X network," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 3827–3847, Feb. 2024.

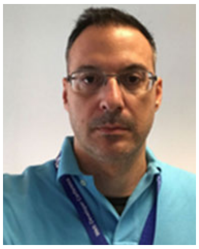
[32] C. R. del-Blanco, P. Carballeira, F. Jaureguizar, and N. García, "Robust people indoor localization with omnidirectional cameras using a grid of spatial-aware classifiers," *Signal Process., Image Commun.*, vol. 93, Apr. 2021, Art. no. 116135.

[33] J. Zhu, J. Zhu, X. Wan, C. Wu, and C. Xu, "Object detection and localization in 3D environment by fusing raw fisheye image and attitude data," *J. Vis. Commun. Image Represent.*, vol. 59, pp. 128–139, Feb. 2019.

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 (CERTH/ITI), 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 ECE 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 (CERTH/ITI). His research interests include 5G/6G networks, V2X communications, artificial intelligence, reconfigurable intelligent surfaces, metamaterials, the IoT in autonomous vehicles, and eHealth domains.

MINAS DASYGENIS (Member, IEEE) 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, in the research area of designing embedded systems and accelerators in homogeneous or heterogeneous architectures. He has over 16 years of teaching experience in operating systems, computer architecture, embedded systems, parallel and distributed systems, and computer networks. His research interests include computer architecture, robotics, embedded and cyber-physical systems, gamification, the Internet of Things, security, and hardware and software cosynthesis.



KONSTANTINOS VOTIS received the Ph.D. degree in computer engineering and informatics, in 2011. Since October 2019, he has been a Visiting Professor with the Institute for the Future, University of Nicosia, regarding blockchain and AI technologies. 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 and personalized healthcare.