IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 June 2024, accepted 4 July 2024, date of publication 8 July 2024, date of current version 26 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3425308

APPLIED RESEARCH

Multimodal Abnormal Event Detection in Public Transportation

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 101077587.

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

ABSTRACT This work addresses the growing concerns about security and passenger safety on public transportation. With the increasing demand for public transport and the rise in road traffic injuries, there is a need for advanced safety measures. Our paper proposes a multi-modal abnormal event detection framework that uses deep learning models in several modalities (RGB, depth, and audio) to identify abnormal events and petty crimes, including passenger aggression, petty thefts, and vandalism. The proposed detection system is designed for use inside autonomous vehicles and, thus, aims to function in the absence of a bus driver. As a result, its main goal is to enhance the safety and security of passengers during transportation. The system methodology involves a deep learning architecture that operates at different framerates and employs multimodal feature extraction and fusion. The experiments for this work were conducted on a custom multi-modal dataset, which includes various classes, including bagsnatch, falldown, fighting, normal, and vandalism, and showcased exceptional detection results compared to other existing action recognition models, with a total accuracy of 85.1%. Finally, the study concludes that the proposed system can be installed in autonomous vehicles and significantly improve safety measures on public transportation.

INDEX TERMS Abnormal event detection, deep learning, embedded systems, edge computing, multimodal, public transportation.

I. INTRODUCTION

Public transportation has become an integral part of people's everyday lives, having a crucial impact on their overall life satisfaction and well-being [1]. However, with the continuous growth of the global population and the respective ever-increasing demand for public transport, security and passenger safety are gradually becoming prominent issues that need to be addressed, as road traffic injuries alone account for approximately 1.3 million deaths each year, according to the World Health Organization (WHO) [2]. As stated in studies from the most recent bibliography, the majority of traffic injuries are caused by human factors like the drivers' habits or their unsafe driving behavior and mainly include fatigue, sleep deprivation [3] and carelessness [4]. To resolve these issues, autonomous vehicles (AVs) were proposed as a way of counteracting injuries caused by driving faults. More specifically, the basic concept behind AVs is the partial or complete replacement of the human factor during driving by mechanical and/or electronic machinery [5]. As a result, enhanced safety and reduced risk of human errors are achieved, while collision accidents caused by human factors are eliminated.

However, this focus on eliminating driver-related risks has overshadowed another critical aspect of passenger safety: the in-cabin environment. Incidents including passenger aggression, simple assaults, petty thefts, and vandalism are some indicative examples of abnormal events that may arise with the absence of an authority figure that could otherwise establish a safe and secure environment for the passengers.

The associate editor coordinating the review of this manuscript and approving it for publication was Hadi Tabatabaee Malazi.

For this purpose, several video surveillance and activity recognition approaches based on deep learning have been proposed for monitoring [6], [7] or identifying abnormal events [8], [9] inside the shuttle. However, visual abnormal event detection is a challenging task to accomplish in computer vision, as the definition alone of an abnormal event is highly dependent on the context.

A large percentage of the actions happening inside the cabin of the autonomous vehicle is almost static. Once aboard, the passengers usually sit on a specific position during their trip with minimal variations, such as movement of hands, as a result of a conversation or slight body variations on vehicle's acceleration/braking due to mass inertia. However, abnormal events are composed of fast motion that occurs in a fragment of a second. Passengers falling down due to a sudden stop in the vehicle's route or a thief stealing a bag and running out of the vehicle are both examples of fast motion. To exploit this large dynamic range of motion, we employ an architecture based on multiple pathways, inspired by [10]. Our framework incorporates both low framerate processing for extracting high-resolution spatial details and high framerate analysis for detecting rapid temporal movements [10], [11]. A slow pathway is designed to capture fine details of a small image sequence while operating in low framerate and low refreshing speed. In parallel, fast pathways with lower channel capacities are focused on fast motion operating in higher framerate and refreshing speed. In our multi-modal approach, we introduce depth information that provide crucial spatial context that 2D images alone cannot capture. This could be particularly useful in video understanding tasks. In addition, audio data can offer additional cues for events that may not be visually apparent. For example, the sound of a breaking glass could indicate an event even if it is not visible in the frame. For this purpose, we design two additional fast pathways, one for the depth modality that captures rapid changes in the spatial configuration of a scene and an additional the acoustic modality in order to capture auditory events. The connections between the different pathways are carefully designed to facilitate meaningful interactions between the modalities. The model was initialized from scratch without using any pre-trained models or transfer learning techniques. We decided against using pre-trained models or transfer learning techniques due to the domain-specific nature of the dataset and the potential for negative transfer. Pre-trained models are typically trained on large-scale datasets with diverse scenarios, which might not align perfectly with the specific context of our dataset. Therefore, we chose to train the model from scratch to ensure that it learns features specific to the task of abnormal event detection in public transportation. However, we recognize the potential benefits of pre-trained models and transfer learning techniques. In future work, we plan to explore the use of these techniques, potentially focusing on domain-specific pre-training or fine-tuning the model with pre-trained weights on similar datasets.

The remainder of this paper is organized in 5 sections as follows. In Section II, related deep learning networks for video recognition are included from the most recent bibliography. In Section III, the methodology of our proposed multi-modal feature extraction framework is introduced for identifying abnormal events in public transportation means, by analyzing the framework architecture, the fusion of the multiple modalities, as well as the data preprocessing steps followed. In Section IV, the experimental results are presented, along with the required hardware components for the system deployment. Finally, Section VI summarizes the conclusions of this study.

II. RELATED WORK

Video recognition has been a focal point of research, evolving from the use of hand-crafted features [12] and shallow learning models [13] to intricate deep learning architectures. Early models like 3D ConvNets [14] and C3D [15] utilized spatiotemporal convolutions, but their capabilities were limited in capturing complex temporal patterns. These early models were foundational but lacked the sophistication to handle the intricacies of video data, which is inherently high-dimensional and requires understanding both spatial and temporal dimensions.

A significant advancement was the SlowFast Network [10], which introduced a dual-pathway architecture to capture both spatial semantics and temporal motion, setting new performance benchmarks. The SlowFast architecture was revolutionary in its approach to disentangling spatial and temporal features, thereby allowing for more efficient and accurate video recognition. However, the SlowFast Networks primarily focused on RGB data, leaving room for improvement in incorporating other modalities.

Another approach, the Two-Stream ConvNets [16], also aimed to capture spatial and temporal features but required training two separate networks, making it computationally expensive. While effective, the Two-Stream ConvNets approach was not as integrated as the SlowFast Network, leading to challenges in real-world applications where computational resources might be limited.

MoViNets [17] have emerged as a mobile-friendly architecture for video understanding, designed to operate efficiently on edge devices. MoViNets have shown promising results in real-time applications but are primarily optimized for computational efficiency rather than recognition performance. This makes them less suitable for applications where high recognition accuracy is required, such as surveillance or medical diagnosis.

For video anomaly detection, various approaches have been thoroughly examined [18] and proposed, including the hierarchical graph embedded pose regularity learning model [19], which utilizes a hierarchical graph structure to capture spatiotemporal dependencies, and the Self-Supervision-Augmented Deep Autoencoder (SSR-AE) [20], which enhances traditional autoencoder methods with

self-supervised tasks to improve feature learning for anomaly detection. In this context, Huang et al. [21] also demonstrated effective model training with limited annotations. Finally, for improving the robustness of anomaly detection by learning a distributional representation of normal events, the AMPNet framework [22] has been proposed, helping in distinguishing anomalous activities effectively.

Various self-supervised approaches have also been proposed, including the Temporal-Aware Contrastive Network (TAC-Net) [23] and the Self-Supervised Attentive Generative Adversarial Network (SSAGAN) [24] frameworks, which leverage contrastive self-supervised learning and combine self-attention mechanisms and self-supervised discriminators to improve anomaly detection. Liu et al. [25] proposed another self-supervised learning approach that enhanced the ability of models to detect anomalies by capturing the distributional properties of normal data, making it easier to identify deviations.

In the realm of multimodal learning [26], several works have explored the use of depth information, such as the work by Izadinia et al. [27], which utilized depth maps for action recognition. Wei et al. [28] proposed a two-stage multi-modal information fusion method for violence detection, using multiple instance learning strategies to refine video-level hard labels into clip-level soft labels, as well as a Multimodal Supervise-Attention enhanced Fusion (MSAF) framework [29] for refining video-level ground truth into pseudo clip-level labels and uses supervise-attention to enhance feature alignment and fusion for predicting anomaly scores. Similarly, the Depth CNN model [30] employed depth information to enhance object detection in videos. These models have shown the utility of depth information in understanding the geometric context of a scene, but they often treat depth as an isolated modality, not fully integrated into the video recognition pipeline.

Approaches for extracting both local spatial features and long-term temporal dependencies from multi-modal driving sequences are also present, demonstrating improved performance and robustness in identifying driving styles. In this context, Liu et al. [31] proposed the DSDCLA framework, which combines a hybrid CNN-LSTM architecture with multi-level attention fusion for driving style detection.

Finally, the role of audio in video recognition has also been explored, with works like those by Owens et al. [32] demonstrating that audio features could provide complementary information in noisy or visually occluded scenarios. However, these works have not been integrated into a unified framework that can handle multiple modalities effectively.

While existing research has made significant contributions to video recognition, particularly in exploring individual modalities like RGB, depth, and audio, there remains a research gap in effectively integrating these modalities into a unified framework for abnormal event detection. Current approaches often treat these modalities as separate channels or supplementary features, limiting their ability to capture the full richness and complexity of real-world events. Moreover, existing multimodal models often come with high computational costs, making them unsuitable for deployment on resource-constrained edge devices often employed in public transportation settings. This work aims to extend the SlowFast architecture by incorporating additional modalities of depth and audio. Unlike existing works that treat these modalities as separate channels or supplementary features, we propose a novel approach that integrates them directly into both the Slow and Fast pathways. This allows our model to capture a more comprehensive representation of videos, encompassing spatial semantics, temporal dynamics, auditory cues, and geometric context. Our unified framework offers a more robust and accurate model that can be deployed on embedded solutions to develop safe and secure services for AVs.

III. METHODOLOGY

Our proposed method can be described as a multi stream architecture that operates at multiple different framerates. We use the concept of pathways to handle the spatiotemporal characteristics of each modality, a slow pathway with a large temporal stride along with multiple faster pathways operating at higher framerate but with lower channel capacity. We employ the slow path only on the RGB modality to extract features with fine details.

A. MULTI-MODAL ARCHITECTURE

In the proposed architecture (Figure 2), the low framerate (LFR) pathway operates at a fraction of the temporal resolution compared to the fast pathways. The slow path uses sampling with a rate of α times slower than the fast pathway. Coupled with a larger channel dimensionality (denoted by β), the slow pathway effectively captures the spatial semantics of the video content. While its temporal granularity is coarser, its design ensures that high-level spatial features are retained, making it a pivotal component for holistic video understanding.

Conversely, the high framerate (HFR) is designed to capture the temporal dynamics within video sequences. Operating at a higher temporal resolution, typically α times faster than the slow pathway, this path ensures fine-grained motion patterns are identified. Although it typically uses a reduced channel dimensionality (often less than the β factor used in the slow pathway), its primary purpose is to grasp short-term temporal dependencies and rapid movements, which, when combined with the spatial insights from the slow pathway, provide a comprehensive video understanding.

Our proposed architecture introduces key modifications to effectively handle multi-modal data. While SlowFast primarily focuses on RGB data, we extend its dual-pathway design to incorporate depth and audio modalities. Specifically, in addition to the slow pathway for RGB, we introduce dedicated fast pathways for depth and audio, enabling the capture of rapid changes in spatial configuration and auditory cues, respectively. This multi-modal approach allows for a

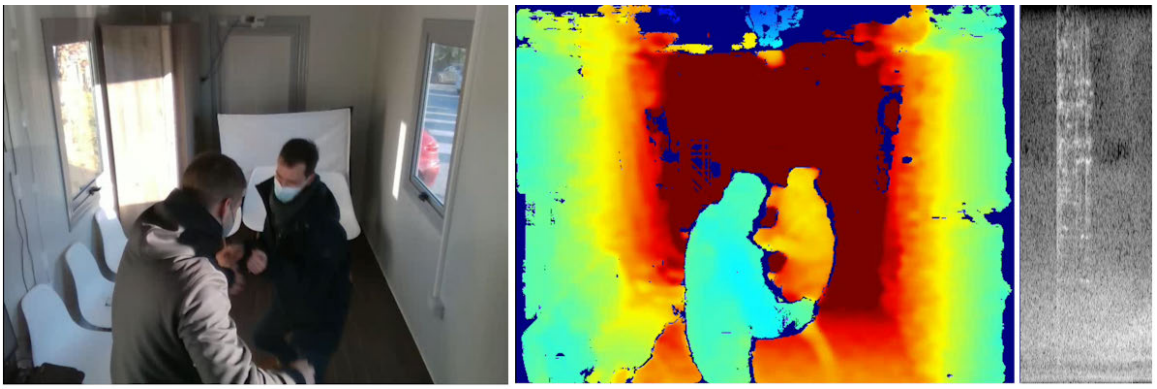


FIGURE 1. Data preprocessing tailored to the (a) RGB modality, (b) depth modality and (c) audio modality. The depth heatmap utilizes a color gradient to represent the distance of objects from the camera.

TABLE 1. Network architecture with ResNet-50 as the backbone, featuring multiple pathways: a slow one with higher channel sizes (denoted by $\beta = 1/8$) and three fast with higher temporal resolution (denoted by $\alpha = 8$) for each modality.

stage	LFR RGB	HFR RGB	HFR Depth	HFR Audio
data layer	stride $16, 1 \times 1$	stride $2, 1 \times 1$	stride $2, 1 \times 1$	stride $2, 1 \times 1$
conv_1	$1 \times 7 \times 7, 64$ stride $1, 2 \times 2$	$5 \times 7 \times 7, 8$ stride $1, 2 \times 2$	$5 \times 7 \times 7, 8$ stride $1, 2 \times 2$	$5 \times 7 \times 7, 8$ stride $1, 2 \times 2$
pool_1	$1 \times 3 \times 3$, maxpooling stride $1, 2 \times 2$	$1 \times 3 \times 3$, maxpooling stride $1, 2 \times 2$	$1 \times 3 \times 3$, maxpooling stride $1, 2 \times 2$	$1 \times 3 \times 3$, maxpooling stride $1, 2 \times 2$
res_2	$\begin{bmatrix} 1 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 1^2, 8 \\ 1 \times 3^2, 8 \\ 1 \times 1^2, 32 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 1^2, 8 \\ 1 \times 3^2, 8 \\ 1 \times 1^2, 32 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 1^2, 8 \\ 1 \times 3^2, 8 \\ 1 \times 1^2, 32 \end{bmatrix} \times 3$
res_3	$\begin{bmatrix} 1 \times 1^2, 128 \\ 1 \times 3^2, 128 \\ 1 \times 1^2, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 1^2, 16 \\ 1 \times 3^2, 16 \\ 1 \times 1^2, 64 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 1^2, 16 \\ 1 \times 3^2, 16 \\ 1 \times 1^2, 64 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 1^2, 16 \\ 1 \times 3^2, 16 \\ 1 \times 1^2, 64 \end{bmatrix} \times 4$
res_4	$\begin{bmatrix} 3 \times 1^2, 256 \\ 1 \times 3^2, 256 \\ 1 \times 1^2, 1024 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 1^2, 32 \\ 1 \times 3^2, 32 \\ 1 \times 1^2, 128 \end{bmatrix} \times 6$	$\begin{bmatrix} 3 \times 1^2, 32 \\ 1 \times 3^2, 32 \\ 1 \times 1^2, 128 \end{bmatrix} \times 6$	$\begin{bmatrix} 3 \times 1^2, 32 \\ 1 \times 3^2, 32 \\ 1 \times 1^2, 128 \end{bmatrix} \times 6$
res_5	$\begin{bmatrix} 3 \times 1^2, 512 \\ 1 \times 3^2, 512 \\ 1 \times 1^2, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$

more comprehensive understanding of the scene compared to using RGB alone.

B. MULTI-MODAL FUSION

The LFR and HFR pathways are fused using lateral connections that are designed to allow the exchange of information between the slow and fast pathways at multiple levels. The reason behind is to ensure that while each pathway focuses on its primary objective (spatial or temporal), it can benefit from the auxiliary information present in the other pathway. Given the feature maps of the fast pathway, F_{fast}, a temporal pooling transformation reshapes the temporal resolution of the slow pathway. This pooled representation is later fused with the feature maps of the slow pathway, F_{slow}. The fusion is achieved using simple operations like sum:

$$F_{\text{merged}} = F_{\text{slow}} + \text{pool}(F_{\text{HFR[color + depth + audio]}}) \quad (1)$$

The fused feature map, F_{merged}, benefits from both the spatial information of F_{slow} and the fast temporal features of F_{fast}. The lateral connections are positioned at different depths, allowing for a rich multi-level exchange of spatiotemporal features, to enhance the capability of the proposed network in recognizing complex video patterns. This design choice is motivated by the need to leverage both the fine-grained spatial information captured by the slow pathway and the temporal dynamics captured by the fast pathways, ensuring a holistic representation of the video.

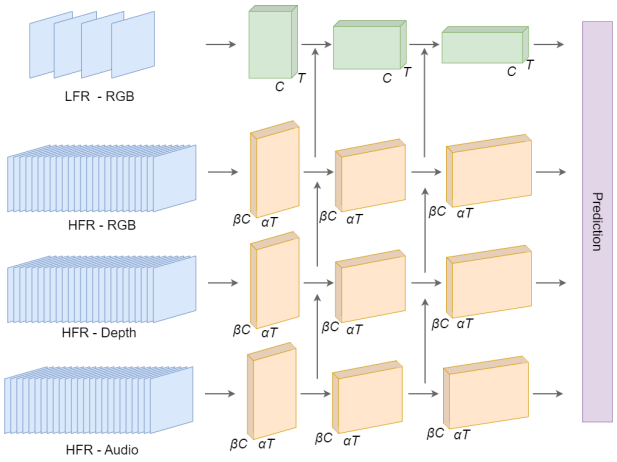


FIGURE 2. General architecture of the proposed multimodal framework for abnormal event detection in public transportation.

C. DATA PREPROCESSING

In the first preprocessing step, individual frames from both modalities were resized to a uniform dimension of 224 × 224 pixels, ensuring consistent input structure and mitigating potential disparities arising from varied frame sizes. Furthermore, normalization was applied to the data using standard deviation as the scaling factor. This step standardized the range of the pixel values, centering them around zero, which is pivotal in enhancing the convergence speed and stability during training. Figure 1(a) illustrates the standard RGB input from the camera feed. The depth heatmap utilizes a color gradient to represent the distance of objects from the camera. Specifically, as seen in Figure 1(b), blue color is used to indicate objects that are far from the camera position, while red shades represent objects that are closer. This color encoding provides an intuitive mapping to utilize the spatial distribution of objects in the scene.

Regarding the acoustic signal, depicted in Figure 1(c), for a 3D magnitude representation, the Mel spectrogram was used as a representation of audio data that extend beyond the traditional 2D spectrograms by adding an additional dimension to capture more information from the audio signal. Specifically, the Mel spectrogram is based on the short-time Fourier transform (STFT), where the audio magnitude is obtained by computing the Fourier transform for successive frames in a signal based on the following equation:

$$X(m, w) = \sum_{n=-\infty}^{+\infty} x(n)w(n-m)e^{-jwn} \tag{2}$$

where n is a discrete time index, m is a discrete window and w is the angular frequency per sample. The $X(m, w)$ is the discrete-time Fourier transform (DTFT) of the signal $x(n)$, which is multiplied by a window function $w(n)$ for a short period of time, providing information about the signal's frequency content as a function of both frequency (w) and the chosen window position (m).

As mentioned above, the Mel spectrogram shares the same representation as the STFT but adjusts the frequency axis to conform to the Mel scale, an approximation of the nonlinear way that frequencies are perceived by individuals, by employing overlapping triangular filters.

IV. EXPERIMENT

In this section, the findings derived from our proposed research methodology are offered with the solely purpose to highlight key observations, trends, as well as insights gained throughout the experimentation process.

A. DATASET

We evaluated our approach on our custom dataset using standard evaluation protocols. Currently, there are no publicly released datasets with aligned data from RGB, depth and acoustic modalities focused on abnormal event detection or action recognition. For this reason, we collected a custom dataset from a simulated environment comprised of video clips, each containing synchronized streams of RGB, depth, and audio data. The dataset was collected for multi-modal learning and informed consent was obtained for every human subject. The categorization of abnormal incidents was established through a collaborative effort involving public transport operators actively participating in the H2020 AVENUE project. We conducted a thorough questionnaire to gather their insights and expertise, drawing upon their extensive experience with police reports and operational data. This collaborative approach ensured that the categorization reflects real-world scenarios and the specific needs of public transport systems. In total, we collected approximately 45 hours of video footage, amounting to 64.2 GB of data. To ensure the dataset's appropriateness for video detection, we took several measures. First, the data includes a wide range of abnormal events to cover different types of dynamic actions, ensuring diverse scenarios. All modalities were synchronized and aligned based on timestamps to provide a comprehensive multimodal dataset. During preprocessing, we applied various data augmentation techniques to enhance the robustness of our model and prevent overfitting. Finally, random sample tests were conducted to ensure that the recorded data was clear, well-lit, and free from excessive noise or artifacts. These steps were taken to ensure that the dataset was suitable for training and evaluating our video detection framework effectively.

For the simulated environment, we used a modular cabin kit and we recruited volunteers to act out the specific scenarios, including bagsnatch, falldown, fighting, vandalism, and normal passenger behavior. Each participant was informed of the purpose of the data collection and provided with a detailed consent form outlining the procedures involved, the data usage, and their rights to withdraw from the study at any time. RGB and depth data were captured using the Orbbec 3D Astra Embedded S sensor and the Intel RealSense D435. Audio was simultaneously recorded using the ReSpeaker 4-Mic Array, equipped with four MEMS microphones,

ensuring temporal alignment across all modalities. The sensors were connected via USB 3.0 interface to the host computer. The host computer is the NVIDIA Jetson AGX Xavier embedded computing platform. It is equipped with an 8-core ARM CPU, a Volta GPU with 512 CUDA cores, and 32 GB of LPDDR4x RAM, offering good computational capabilities in a compact form factor with low energy consumption. With a peak performance of up to 32 TOPS (Tera Operations Per Second) and a power envelope ranging from 10W to 30W, the Jetson AGX Xavier is cost-efficient platform for our use case. The simulated scenarios were designed to mimic real-world events as closely as possible. For instance, the bagsnatch scenario involved an actor simulating a thief snatching a bag from another actor, while the falldown scenario involved an actor simulating a passenger falling due to a sudden stop. These scenarios were carefully designed to capture the characteristics of different abnormal events, including the speed of motion, the interaction between actors, and the overall context of the events. The dataset was split into training, validation, and test sets, with 80% of the data used for training, 10% for validation, and 10% for testing. The split was both random and stratified to ensure balanced representation of all classes across the different sets. This ensures that each class is proportionally represented, thereby enhancing the reliability and robustness of the model's performance evaluation.

TABLE 2. Statistics of the custom multi-modal dataset.

Class	Size (GB)	Samples	Duration	FPS
bagsnatch	9.8	522	2h 10m 30s	15
falldown	9.1	467	1h 56m 45s	15
fighting	13.4	662	2h 45m 30s	15
normal	19.3	990	4h 7m 30s	5-15
vandalism	12.6	623	2h 35m 45s	15

B. TRAINING

We trained our model with our multi-modal dataset, which is initialized from scratch without the use of any pre-trained models. The dataset is comprised of various classes including bagsnatch, falldown, fighting, normal and vandalism as illustrated in Table 2. We conduct training over 250 epochs, which is sufficient for effective learning given the dataset's complexity and size. The models employ a momentum setting of 0.9 and a weight decay factor of 10^{-4}. To mitigate the risk of overfitting, a dropout rate of 0.5 is applied before the final classification layer. A custom data augmentation generator dynamically alters the training samples during each epoch, incorporating a variety of transformations, such as random rotations, zooming, positional shifts, and horizontal flipping. This approach not only enriches our dataset without requiring additional storage, but also plays a crucial role in improving the model's predictive performance and adaptability to a wide array of new and diverse data. By consistently exposing the model to these transformed instances, we ensure that it learns features that are generalized rather than simply memorizing the training set. Figure 3 depicts the improvement in validation accuracy with the additional modalities.

Figure 4 presents the validation accuracy achieved at various epochs for three different modalities: RGB, RGBD (RGB+Depth), and RGBDA (RGB+Depth+Audio). RGB represents the use of only color information, RGBD incorporates both color and depth, and RGBDA includes color, depth, and audio. As the training progresses, there is a noticeable improvement in validation accuracy for all modalities. By epoch 250, the RGBDA modality achieves the highest validation accuracy of 85.1%, followed by RGBD at 82.6%, and RGB at 77.5%. This suggests that the incorporation of additional modalities like depth and audio significantly enhances the model's performance, with RGBDA showing the most promising results.

C. MAIN RESULTS

As illustrated in Table 3, the proposed approach demonstrates a high degree of accuracy, with Top-1 Accuracy ranging from 83.0% for the "falldown" class to 87.6% for the "normal" class. The F1-Scores, which are harmonic means of Precision and Recall, also confirm the model's robustness, falling within the range of 0.828 to 0.875. Precision and Recall metrics further validate the model's ability to correctly identify and classify activities, with values consistently above 0.8 across all classes. One notable observation is that the model sometimes confuses the classes of fighting and bagsnatch due to their similar dynamic movements. Both involve rapid and abrupt actions that can be challenging to distinguish without more contextual information. This highlights an area for potential improvement in future iterations of our model. The Support column indicates the number of test instances for each class in the dataset, providing context for the performance metrics. The table also includes an aggregate metric, showing a total Top-1 Accuracy of 85.1% across all 652 test instances. This suggests that the proposed model performs well not just in isolated classes but also when evaluated on the dataset as a whole. Overall, the results indicate that the model is both precise and reliable, making it a strong candidate for real-world applications in activity recognition.

TABLE 3. Class-wise performance Metrics for the proposed approach.

Class	Top-1 (%)	F1-Score	Precision	Recall	Support
Bagsnatch	84.1	0.840	0.852	0.829	104
Falldown	83.0	0.828	0.839	0.818	93
Fighting	83.8	0.836	0.848	0.825	132
Normal	87.6	0.875	0.884	0.867	198
Vandalism	84.6	0.844	0.855	0.834	125
Total	**85.1**	–	–	–	652

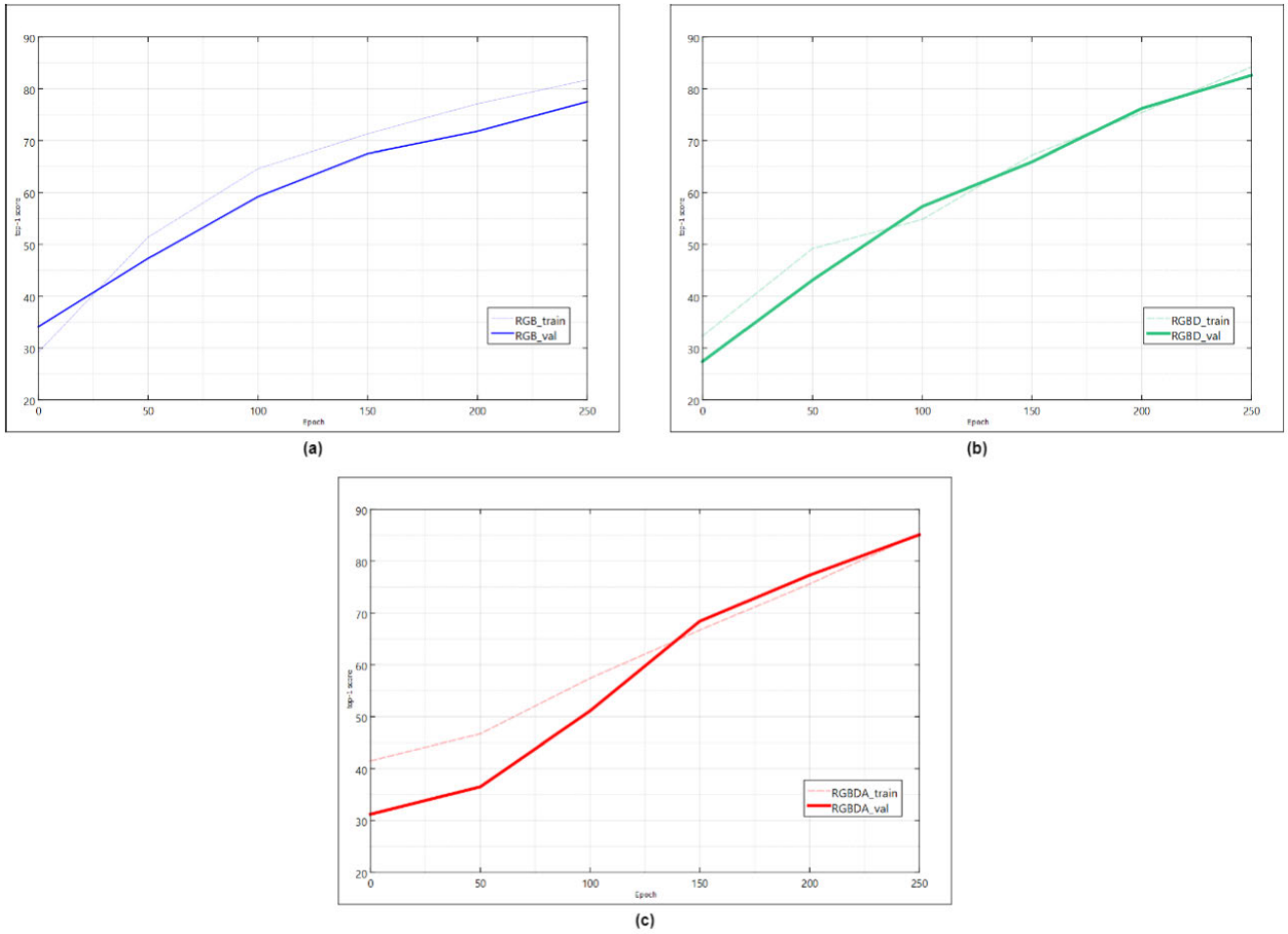


FIGURE 3. Training and validation results of the proposed deep learning framework on: (a) RGB modality; (b) RGB + Depth modalities; and (c) RGB + Depth + Audio modalities. The RGBDA model outperforms others with the highest accuracy and the smallest gap between training and validation accuracy, suggesting its better generalization capability. The results indicate that incorporating more data modalities improves model performance for the given task.

To gain insights into the regions of the input that are most influential for predictions, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) [33]. This technique provides a high-resolution, class-discriminative visualization by using the gradients of the target class flowing into the final convolutional layer of our model. By overlaying these heatmaps onto the original input (Figure 5), we can visually interpret which areas are pivotal for the model's decision-making process. This is particularly useful for understanding the model's behavior across the various classes in our custom multi-modal dataset, such as bagsnatch, falldown, fighting, normal, and vandalism. The use of Grad-CAM not only aids in model interpretability, but also serves as a diagnostic tool for identifying potential areas where the model might be focusing incorrectly, thereby guiding further refinements in the training process. As illustrated in Table 4, the proposed model, which incorporates multiple modalities (RGBDA), demonstrates superior performance with a top-1 accuracy of 85.1% on our custom dataset. Interestingly, MoviNet-A0 stands out for its efficiency, requiring only 6.1 GFLOPs, while still achieving a respectable accuracy of 75.2%. On the other hand, MoviNet-A6 offers the highest accuracy of 81.5% but at a higher computational cost of 117 GFLOPs. SlowFast and X3D-XL offer balanced performances but are outperformed by the proposed model in terms of accuracy. Overall, the table serves as a valuable reference for understanding the trade-offs between accuracy, computational complexity, and model size in the domain of action recognition. The proposed architecture aims to leverage multi-modal data for enhanced video understanding. By incorporating depth and audio modalities, the model could potentially capture a richer set of features that are not apparent in RGB data alone. This could be particularly useful in complex scenarios where visual data might be ambiguous or insufficient for accurate classification or recognition.

However, the addition of multiple fast paths could increase the computational cost, which would need to be carefully evaluated against the performance gains. Overall, this research direction could offer valuable insights into the

TABLE 4. Comparison of action recognition models.

Feature/Algorithm	SlowFast	MoviNet-A0	MoviNet-A6	X3D-XL	I3D	C3D	**ours**
Architecture	Dual-pathway	Stream-buffer	Stream-buffer	3D Expanded	Inflated 2D	3D ConvNet	Multi-pathway
Backbone	ResNet-50	A0	A6	ResNet	Inflated Inception-V1	-	ResNet-50
Modalities	RGB	RGB	RGB	RGB	RGB	RGB	Multimodal RGBDA
Resolution	224 × 224	224 × 224	224 × 224	224 × 224	224 × 224	112 × 112	224 × 224
Parameter Count (M)	35	5.3	32	12	12.3	78	47
GFLOPs	36.1	**6.1**	117	48.4	55.2	38	48.2
Top-1 Accuracy (%)	79.4	75.2	81.5	77.4	72.6	68.2	**85.1**

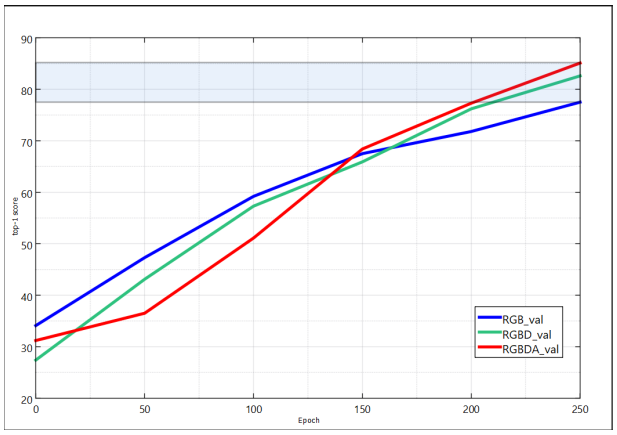


FIGURE 4. Validation accuracy achieved at various epochs for three different modalities: RGB, RGB + Depth, and RGB + Depth + Audio.

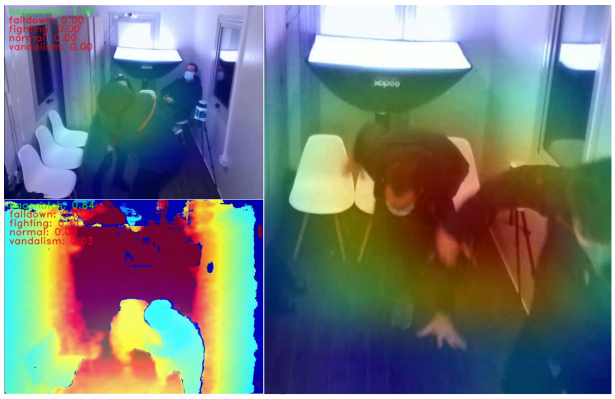


FIGURE 5. Visualization of Grad-CAM activations overlaid on original input samples. The heatmaps highlight the regions that are most influential in the model's decision-making process for different classes.

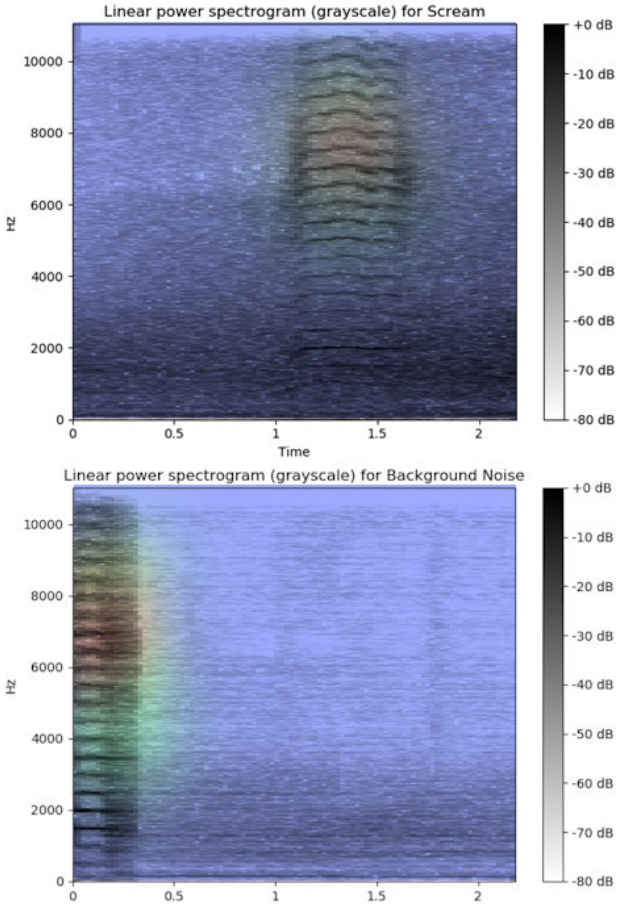


FIGURE 6. Activation maps on Mel-Spectrograms: (Top) Correctly classified screaming event. (Bottom) False positive detection of background noise as screaming due to microphone distortion.

effective fusion of multi-modal data in video recognition tasks.

D. OPTIMIZATIONS

The deployment of deep learning models on edge devices, is challenged due to the various hardware and infrastructure constraints. While edge devices offer the advantage of localized computation, reducing latency and bandwidth usage, they also come with limitations in terms of computational resources and power consumption. To address these challenges, we use a series of performance optimizations, aiming to achieve real-time video action recognition on the Jetson AGX Xavier platform. TensorRT framework optimizations using NVIDIA tools were applied for model quantization, offloading to hardware DLA (Deep Learning Accelerators) and layer fusion. These optimizations are performed automatically using 'trtexec' and effectively

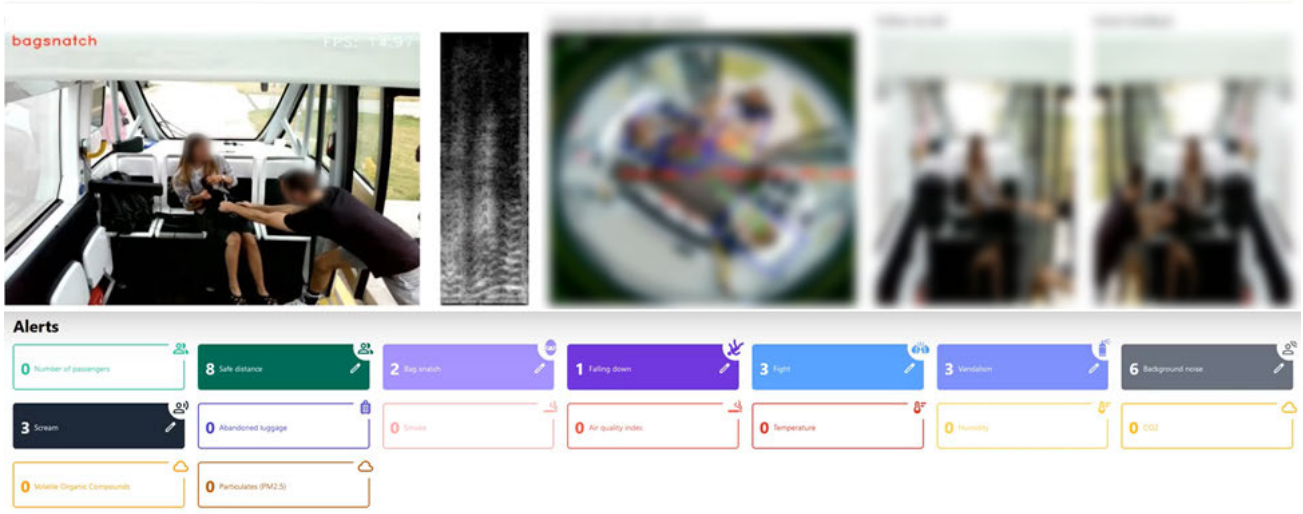


FIGURE 7. Alerts visualized and metadata on the operator's dashboard regarding the detected incidents.

reduce the model's computational footprint, allowing it to run efficiently within the limited resources of the edge device. Quantization reduces the precision of the model's weights and activations, decreasing memory footprint and improving inference speed. DLA offloading leverages the specialized hardware on the Jetson AGX Xavier to accelerate convolutional neural network operations, further enhancing performance. Finally, layer fusion combines multiple layers into a single operation, streamlining the computation graph and reducing memory access overhead. The procedure included the conversion to ONNX format and the TRT engine generation using the 'trtexec' tool. To validate the effectiveness of our optimizations, we conducted extensive benchmarking. As shown in Table 5, the optimized model achieved an 2.5x speedup compared to the unoptimized version, with a small drop of 0.5% in Top-1 Accuracy. These results demonstrate the efficacy of our optimization strategies and confirm that the proposed model, when appropriately optimized, is well-suited for real-time video action recognition on resource-constrained devices.

E. EVALUATION ON NAVYA VEHICLES

The proposed system was evaluated in automated minibuses in Copenhagen and Geneva (Figure 8). During the validation, fighting, bag snatch, falling down and vandalism scenarios were detected in vehicle P109 that serves the route Slagelse in Denmark. The results illustrated in Table 6 indicated that the algorithms were able to correctly identify most of the performed scenarios with 89% accuracy and the appropriate notifications were also captured in the operator's dashboard (Figure 7). The system exhibits an average response time of approximately 1200 ms. This is attributed to the fact that our approach operates on a sliding buffer of consecutive frames to capture temporal information effectively.

TABLE 5. Performance optimizations applied on the Jetson platform.

Optimization Technique	Speedup
INT8 Quantization	1.5x
DLA Offloading	1.3x
Layer Fusion	1.2x
Asynchronous DataLoader	1.1x
GPU-based Image Trasformation	1.1x
Cumulative Effect	**2.8x**

TABLE 6. On-site evaluation metrics for each class.

Abnormal Event	Accuracy	F1-Score	Samples
Bagsnatch	0.90	0.91	5
Falldown	0.89	0.89	4
Fighting	0.83	0.86	6
Vandalism	0.92	0.91	5

V. DISCUSSION

It is important to note that currently, there is a lack of publicly available datasets that include synchronized and aligned RGB, depth, and audio data specifically tailored for abnormal event detection or action recognition in public transportation settings. This limitation makes it challenging to conduct a direct comparison with existing methods that primarily focus on RGB data. However, we believe that our proposed multi-modal approach offers significant advantages in capturing a richer set of features, leading to improved accuracy and robustness compared to relying solely on visual information. In the future, we aim to contribute to the research community by creating and releasing such a multi-modal dataset, enabling a more comprehensive evaluation and comparison of different approaches in this domain.

FIGURE 8. Installation of the system in vehicle. The green bounding box highlights the Jetson device, powered through the vehicle's battery and interconnected with the various sensors such as the stereo camera and microphone array (enclosed in the red bounding box).

FIGURE 9. Scenarios with passengers: bag snatching, fighting, vandalism, and falling down. Red overlay indicates the activation maps as visual feedback, demonstrating the focus points of the algorithm during real-time operation within the pilot site environment.

Moreover, a detailed ablation study will be conducted to assess the contribution of each modality to the overall performance of the network.

VI. CONCLUSION

This study introduces an advanced multi-modal abnormal event detection system designed to enhance public transportation safety. The proposed system employs a novel architecture that integrates RGB, depth, and audio data to monitor and identify abnormal events in real time. Experimental results indicate that the system achieves a high accuracy of 85.1%, surpassing existing methods which rely on single-modality data.

The research makes several key contributions. It presents a robust framework that combines multiple data modalities to improve event detection accuracy. This multi-pathway architecture captures both spatial and temporal features, providing a comprehensive understanding of events. The proposed system also demonstrates the feasibility of real-time, in-cabin monitoring, crucial for enhancing passenger safety in autonomous vehicles and other public transportation systems. Practical validation through deployment in autonomous minibuses showcases the system's applicability in real-world scenarios and its efficiency on edge devices like the NVIDIA Jetson.

Despite these advancements, the study has certain limitations. The custom dataset is captured from simulated environments and may not grasp the full spectrum of real-world complexity. The integration of multiple modalities increases computational requirements, and although optimizations have been applied, further enhancements are necessary to ensure the efficiency on resource-constrained devices. Additionally, the system sometimes confuses similar dynamic events, such as fighting and bagsnatching, indicating a need for more sophisticated feature extraction and contextual analysis.

Future research should focus on expanding real-world data collection from diverse public transportation environments to enhance the system's robustness and generalizability. Conducting detailed ablation studies to assess the individual contributions of each modality will provide insights and help optimizing the system's architecture. Exploring the integration of this abnormal event detection system with other public transportation safety systems, such as emergency response and communication protocols, can further enhance overall passenger safety and system efficiency.

In conclusion, this research sets a new standard for real-time abnormal event detection in public transportation, addressing critical safety concerns and contributing valuable insights to the fields of deep learning and computer vision.

REFERENCES

[1] L. E. Olsson, T. Gärling, D. Ettema, M. Friman, and S. Fujii, "Happiness and satisfaction with work commute," *Social Indicators Res.*, vol. 111, no. 1, pp. 255–263, Mar. 2013.
[2] *Road Traffic Injuries*. Accessed: Jul. 7, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
[3] S. Kwon, H. Kim, G. S. Kim, and E. Cho, "Fatigue and poor sleep are associated with driving risk among Korean occupational drivers," *J. Transp. Health*, vol. 14, Sep. 2019, Art. no. 100572.
[4] D. Q. Nguyen-Phuoc, O. Oviedo-Trespalacios, T. Nguyen, and D. N. Su, "The effects of unhealthy lifestyle behaviours on risky riding behaviours—A study on app-based motorcycle taxi riders in Vietnam," *J. Transp. Health*, vol. 16, Mar. 2020, Art. no. 100666.
[5] S. E. Shladover, "Connected and automated vehicle systems: Introduction and overview," *J. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 190–200, May 2018.
[6] D. Tsiktsiris, A. Lalas, M. Dasygenis, K. Votis, and D. Tzovaras, "An efficient method for addressing COVID-19 proximity related issues in autonomous shuttles public transportation," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.*, Hersonissos, Greece. Cham, Switzerland: Springer, Jun. 2022, pp. 170–179.
[7] D. Tsiktsiris, A. Lalas, M. Dasygenis, K. Votis, and D. Tzovaras, "Enhanced security framework for enabling facial recognition in autonomous shuttles public transportation during COVID-19," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.*, Hersonissos, Greece. Cham, Switzerland: Springer, Jun. 2021, pp. 145–154.

[8] 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.

[9] D. Tsiktsiris, A. Vafeiadis, A. Lalas, M. Dasygenis, K. Votis, and D. Tzovaras, "A novel image and audio-based artificial intelligence service for security applications in autonomous vehicles," *Transp. Res. Proc.*, vol. 62, pp. 294–301, Jan. 2022.

[10] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.

[11] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR05)*, Jun. 2005, pp. 886–893.

[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[14] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 568–576.

[17] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "MoViNets: Mobile video networks for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16015–16025.

[18] Y. Liu, D. Yang, Y. Wang, J. Liu, J. Liu, A. Boukerche, P. Sun, and L. Song, "Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–38, Jul. 2024.

[19] C. Huang, Y. Liu, Z. Zhang, C. Liu, J. Wen, Y. Xu, and Y. Wang, "Hierarchical graph embedded pose regularity learning via spatio-temporal transformer for abnormal behavior detection," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 307–315.

[20] C. Huang, Z. Yang, J. Wen, Y. Xu, Q. Jiang, J. Yang, and Y. Wang, "Self-supervision-augmented deep autoencoder for unsupervised visual anomaly detection," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13834–13847, Dec. 2022.

[21] C. Huang, C. Liu, J. Wen, L. Wu, Y. Xu, Q. Jiang, and Y. Wang, "Weakly supervised video anomaly detection via self-guided temporal discriminative transformer," *IEEE Trans. Cybern.*, vol. 54, no. 5, pp. 3197–3210, May 2024.

[22] Y. Liu, J. Liu, K. Yang, B. Ju, S. Liu, Y. Wang, D. Yang, P. Sun, and L. Song, "AMP-net: Appearance-motion prototype network assisted automatic video anomaly detection system," *IEEE Trans. Ind. Informat.*, vol. 20, no. 2, pp. 2843–2855, Feb. 2024.

[23] C. Huang, Z. Wu, J. Wen, Y. Xu, Q. Jiang, and Y. Wang, "Abnormal event detection using deep contrastive learning for intelligent video surveillance system," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5171–5179, Aug. 2022.

[24] C. Huang, J. Wen, Y. Xu, Q. Jiang, J. Yang, Y. Wang, and D. Zhang, "Self-supervised attentive generative adversarial networks for video anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 9389–9403, Nov. 2023.

[25] J. Liu, Y. Liu, W. Zhu, X. Zhu, and L. Song, "Distributional and spatial–temporal robust representation learning for transportation activity recognition," *Pattern Recognit.*, vol. 140, Aug. 2023, Art. no. 109568.

[26] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[27] H. Izadinia, Q. Shan, and S. M. Seitz, "IM2CAD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2422–2431.

[28] D.-L. Wei, C.-G. Liu, Y. Liu, J. Liu, X.-G. Zhu, and X.-H. Zeng, "Look, listen and pay more attention: Fusing multi-modal information for video violence detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1980–1984.

[29] D. Wei, Y. Liu, X. Zhu, J. Liu, and X. Zeng, "MSAF: Multimodal supervise-attention enhanced fusion for video anomaly detection," *IEEE Signal Process. Lett.*, vol. 29, pp. 2178–2182, 2022.

[30] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2015, pp. 681–687.

[31] J. Liu, Y. Liu, D. Li, H. Wang, X. Huang, and L. Song, "DSDCLA: Driving style detection via hybrid CNN-LSTM with multi-level attention fusion," *Appl. Intell.*, vol. 53, no. 16, pp. 19237–19254, Aug. 2023.

[32] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 801–816.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[34] D. Stanojević, P. Stanojević, D. Jovanović, and K. Lipovac, "Impact of riders' lifestyle on their risky behavior and road traffic accident risk," *J. Transp. Saf. Secur.*, vol. 12, no. 3, pp. 400–418, Mar. 2020.

[35] E. Okur, S. H. Kumar, S. Sahay, and L. Nachman, "Multimodal understanding of passenger intents in autonomous vehicles," Dec. 2019, doi: 10.13140/RG.2.2.24143.25762.

[36] Q. Portes, J. Pinquier, F. Lerasle, and J. M. Carvalho, "Multimodal human interaction analysis in vehicle cockpit," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 2118–2124.

[37] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Computer Vision—ECCV*, Crete, Greece. Cham, Switzerland: Springer, 2010, pp. 140–153.

[38] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[39] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580*.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[41] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 20–36.

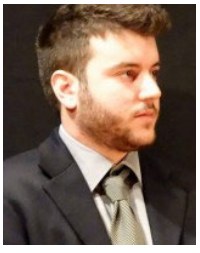
[42] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6816–6826.

[43] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. ICML*, 2021, vol. 2, no. 3, p. 4.

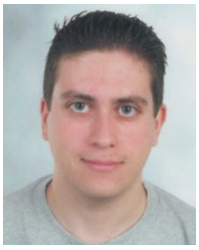
[44] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 200–210.

[45] Y. Li, Z. Lu, X. Xiong, and J. Huang, "PERF-net: Pose empowered RGB-flow net," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 798–807.

[46] Z. Qiu, T. Yao, C.-W. Ngo, X. Tian, and T. Mei, "Learning spatio-temporal representation with local and global diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12048–12057.



DIMITRIS TSIKTSIRIS received the Diploma degree in informatics and telecommunications engineering from the Faculty of Engineering, University of Western Macedonia, in 2017, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. He has been a Research Assistant with the Centre for Research and Technology Hellas, Informatics and Technology Institute (CERTH/ITI), since September 2019. His primary 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 2006. He is currently a Postdoctoral Researcher with the Centre for Research and Technology, Hellas/Information Technologies Institute (CERTH/ITI). From 2020 to 2021, he was an Adjunct Lecturer with the Department of Electrical and Computer Engineering, University of Western Macedonia (UOWM), where he was an Adjunct Lecturer with the Department of Informatics and Telecommunications Engineering, from 2012 to 2018. He was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, AUTH, from 2013 to 2015. 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 (Member, IEEE) received the Ph.D. degree in electrical and computer engineering, in 1999. 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 carries 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 co-synthesis. He is a System Architect in embedded systems and ICT, has been serving as a program committee member or a reviewer to various flagship conferences of embedded systems, has published more than 85 papers in international journals and conferences, authored three books, and he has been a principal researcher in three European research projects.

KONSTANTINOS VOTIS received the Ph.D. degree in electrical and computer engineering, in 2002. He is currently a Computer Engineer and a Senior Researcher (Researcher Grade B') with the Centre for Research and Technologies Hellas, Information Technologies Institute (CERTH/ITI), and the Director of the Visual Analytics Laboratory, CERTH/ITI. He has been a Visiting Professor with the Institute of the Future, University of Nicosia, regarding blockchain and AI technologies, since October 2019. 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.